Trustworthiness and Expertise: Social Choice and Logic-based Perspectives

A thesis submitted in partial fulfilment of the requirement for the degree of Doctor of Philosophy

Joseph Singleton

November 2022

Abstract

This thesis studies problems involving unreliable information. We look at how to aggregate conflicting reports from multiple unreliable sources, how to assess the trustworthiness and expertise of sources, and investigate the extent to which the truth can be found with imperfect information. We take a formal approach, developing mathematical frameworks in which these problems can be formulated precisely and their properties studied. The results are of a conceptual and technical nature, which aim to elucidate interesting properties of the problem at the core abstract level.

In the first half we adopt the axiomatic approach of *social choice theory*. We formulate *truth discovery* – the problem of aggregating reports to estimate true information and reliability of the sources – as a social choice problem. We apply the axiomatic method to investigate desirable properties of such aggregation methods, and analyse a specific truth discovery method from the literature. We go on to study ranking methods for *bipartite tournaments*. This setting can be applied to rank sources according to their accuracy on a number of topics, and is also of independent interest.

In the second half we take a logic-based perspective. We use modal logic to formalise the notion of expertise, and explore connections with knowledge and truthfulness of information. We use this as the foundation for a belief change problem, in which reports must be aggregated to form beliefs about the true state of the world and the expertise of the sources. We again take an axiomatic approach – this time in the tradition of belief revision – where several postulates are proposed as rationality criteria. Finally, we address *truth-tracking*: the problem of finding the truth given non-expert reports. Adapting recent work combining logic with formal learning theory, we investigate the extent to which truth-tracking is possible, and how truth-tracking interacts with rationality.

Contents

List of Figures

List of Tables

List of Publications

The content of this thesis is derived from the following publications.

Chapter 2.

- Joseph Singleton and Richard Booth (2020). "An Axiomatic Approach to Truth Discovery". In: *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '20. Virtual Event: International Foundation for Autonomous Agents and Multiagent Systems, pp. 2011–2013. URL: https://www.ifaamas.org/Proceedings/aamas2020/pdfs/p2011.pdf
- Joseph Singleton and Richard Booth (2022a). "Towards an axiomatic approach to truth discovery". In: *Autonomous Agents and Multi-Agent Systems* 36.2, pp. 1–49. URL: https://doi.org/10.1007/s10458-022-09569-3

Chapter 3.

- Joseph Singleton and Richard Booth (2021). "Rankings for Bipartite Tournaments via Chain Editing". In: *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '21. Virtual Event: International Foundation for Autonomous Agents and Multiagent Systems, pp. 1236–1244. URL: https://www.ifaamas.org/Proceedings/aamas2021/pdfs/p1236.pdf

Chapter 4.

- Joseph Singleton (2021). "A Logic of Expertise". In: *ESSLLI 2021 Student Session*. URL: https://arxiv.org/abs/2107.10832
- Joseph Singleton and Richard Booth (2023). "Expertise and information: an epistemic logic perspective". In: *Synthese* 201.2. URL: https://doi.org/10.1007/s11229-023-04064-y

Chapter 5.

- Joseph Singleton and Richard Booth (2022b). "Who's the Expert? On Multi-source Belief Change". In: *Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning*, pp. 331–340. URL: https://doi.org/10.24963/kr.2022/33

Chapter 6.

- Joseph Singleton and Richard Booth (2022c). "Truth-Tracking with Non-Expert Information Sources". In: *Proceedings of the 20th International Workshop on Non-Monotonic Reasoning*. NMR '22. CEUR Workshop Proceedings, pp. 80–91. URL: http://ceur-ws.org/Vol-3197/paper8.pdf

Acknowledgements

I would like to first thank my supervisor Dr. Richard Booth for his enthusiasm and guidance throughout this project. It's been a real pleasure to work together. I also thank Andreas Herzig and Steven Schockaert for examining the thesis and for their insightful questions. I thank my parents Lucy and Stephen for their continued support, and my brother Ivan for helping me practise my conference talks. I'd like to also thank my fellow students at Cardiff – especially my office-mates Aric and Mohammed – for their company as I wrote this thesis. Finally, thanks to Scott for being here throughout.

— 1 —
Introduction

The overall theme of this thesis is *unreliable information*. How should unreliable information be aggregated? Who should be trusted when conflicts arise between unreliable sources? And what, if anything, can be *learned* from non-expert information? These are the central issues the thesis aims to address.

Indeed, methods for understanding and reasoning with unreliable information are becoming ever more relevant in today's world, as the volume of data produced and consumed grows year-on-year. With the growth of user-generated content on the internet, most prominently on social media platforms, false information can spread rapidly – sometimes with dramatic consequences. As such, much research effort has gone into identifying false information, estimating source reliability, and understanding the nature of how people may come to believe and share false information.

The gap this thesis aims to fill concerns *formal models* of problems surrounding unreliable information. We take a mathematical approach, putting forward formal frameworks in which the relevant problems can be formulated precisely. In doing so we obtain conceptual results which aim to shine light on the core, abstract features of such problems. It is hoped that the thesis complements the various empirical, practical and philosophical approaches to our topic, and contributes to the broader understanding of trustworthiness, expertise and unreliable information from a mathematical point of view.

The thesis is split into two parts along methodological lines. In Chapters 2 and 3 we use the tools and ideas of *computational social choice theory* (Brandt, Conitzer, et al. 2016b) to explore the problems of aggregating unreliable information and ranking sources by trustworthiness. Chapters 4 to 6 take a *logic-based* approach. We develop a modal logic framework to give precise semantics for expertise, and explore the connection between expertise, knowledge, and the truthfulness of information. This framework is used as the foundation for a belief change problem in the tradition of knowledge representation and rational belief change (Booth and Meyer 2011; Hansson 2022; Fermé and Hansson 2018). Finally, we combine ideas from *formal learning theory* (Jain et al. 1999; Gierasimczuk 2010, §2.1) (and in particular, the intersection of formal learning theory and belief revision (Baltag, Gierasimczuk, and Smets 2019)) to investigate the extent to which one can *learn* from unreliable information. In the remainder of this chapter we briefly survey the literature for both halves.

1.1 Social Choice Perspectives

Broadly speaking, social choice theory is the study of *aggregating preferences*. The prototypical example is *voting*. In an election, each member of the electorate submits a vote in the form of their preferences over the candidates. A voting rule then aggregates these preferences into a *collective decision* by declaring the winning candidate, the runner-up, and so on. There are, of course, many different voting methods which can be used to aggregate votes, and several distinct methods are in use in different contexts across the world.

In analysing and comparing such methods, the *axiomatic approach* has been a crucial methodological tool since the seminal work of Arrow (1952), who initiated the age of so-called *classical social choice theory*. In taking this approach, one formalises intuitively desirable properties called *axioms*, which are expected to hold for "reasonable" aggregation methods. This provides a *normative* basis on which to construct and judge voting rules. The benefits of this approach were already shown by Arrow, who proved, surprisingly, that it is mathematically impossible for any voting method to simultaneously satisfy a short list of seemingly desirable axioms. This type of result – known as an *impossibility theorem* – highlights a fundamental and inescapable property of voting.[1] This has practical consequences: if one needs to actually choose a rule for use in a vote, which axiom will be sacrificed?

Axiomatic analysis can also be applied in a *descriptive* context, where one starts with a known voting rule and aims to find axioms satisfied by it. In many cases a set of axioms can be found to characterise a particular rule completely, in that it is the *unique* rule satisfying them. This gives additional insight into the nature of the rule and in how different axioms interact.

The axiomatic approach has since been adapted to various domains besides voting, including tournaments (Brandt, Brill, et al. 2016), judgement aggregation (Endriss 2016), the ranking of web pages (Altman and Tennenholtz 2005), reputation systems (Tennenholtz 2004) and collective annotation (Kruger et al. 2014). While the aggregation problems of each domain have their own unique characteristics, some "standard" axioms appear across the board. While the precise mathematical formulation of such axioms varies across problems, the general intent is the same. We use the example of voting to illustrate a few. In this setting each voter submits a ranking over the set of candidates (the voter's *ballot*), which a voting rule aggregates to form a collective ranking.

> **Anonymity**. All voters are treated equally: if voters i and j swap their ballots, the collective ranking remains the same.

[1] At least, a property of voting in the sense of Arrow's formal framework.

Neutrality. All candidates are treated equally: if all voters swap the positions of candidates c and d in their ballots, the positions of c and d are also swapped in the collective ranking.

Pareto Optimality. If all voters rank candidate c strictly above d, then so too does the collective ranking.

Independence of Irrelevant Alternatives (IIA). The relative ranking of any two candidates c and d in the collective ranking depends only on the rankings of c and d in each voter's ballot.

Positive Responsiveness. If c ranks above d in the collective ranking and some voter changes their ballot to rank c above d, then this remains so in the collective ranking.

Some of these axioms are more clearly desirable than others. **Anonymity** and **Neutrality** are fairly straightforward fairness requirements,[2] and **Pareto Optimality** is generally seen as uncontroversial. **IIA**, however, has come to be seen as a deceptively strong requirement, and plays a role in Arrow's impossibility theorem.

In the last 20 years, *computational social choice* (Brandt, Conitzer, et al. 2016a) has combined social choice theory with ideas from theoretical computer science. For example, complexity theory has been used to show certain voting rules are resistant to strategic manipulation (Conitzer and Walsh 2016), approximation algorithms have been developed for computationally difficult rules (Brandt, Brill, et al. 2016), and SAT solvers have been used to automatically discover new impossibility theorems (Endriss 2020).

In the spirit of combining the axiomatic approach with computer science, we proceed in Chapters 2 and 3 by introducing a social-choice-style framework and several axioms for the problems of *truth discovery* and *bipartite tournament ranking*.

Truth discovery. Truth discovery has arisen recently as a branch of the literature on crowdsourcing (Y. Li, Gao, et al. 2016). When dealing with crowdsourced data one has no guarantees on its veracity: crowdsourcing workers are not generally experts, may exhibit biases, and may even maliciously provide false information. False information provided in this way leads to conflicts among workers, and the dual goals of truth discovery: to find the *true information* in light of conflicts, and to identify the *trustworthy information sources*. The key principle underlying truth discovery methods is the *mutual dependence* between these two goals. According to this principle, sources that provide true information are likely to be trustworthy, and information from trustworthy sources is likely to be true. When sources provide information about multiple objects of interest (e.g. images to be classified), the patterns

[2] Although there are situations where even these do not hold; see Zwicker (2016, p. 32) for examples.

of agreement and disagreement can be used in conjunction with this principle to estimate both truth and trustworthiness.

Most truth discovery work in the literature has focussed on introducing new methods and evaluating them empirically.[3] Our aim in Chapter 2 is to instead study properties of truth discovery methods more generally, by emphasising the similarities between truth discovery and social choice aggregation problems. To that end, we introduce a mathematical framework in which social-choice-style axioms can be expressed. We then explore the interplay between the axioms, and analyse a particular well-known method from the literature.

Bipartite tournaments. While many truth discovery methods in the literature focus on the unsupervised case, so-called *semi-supervised truth discovery* – in which one has access to a subset of ground truth data – has also been studied (Yin and Tan 2011; Rekatsinas et al. 2017). In this situation it is known when sources were correct on the ground truth objects, and this information feeds into the truth discovery process. Using this ground truth may not be straightforward, however, if objects vary in their *difficulty*. For example, it may be preferable to reward sources for correct answers on difficult objects, or penalise them for failures on easy objects. Moreover, difficulty may not be known *a priori*, and is itself subject to estimation.

In Chapter 3 we observe and generalise the key features of this problem: we have entities of two different kinds (information sources and objects of interest), comparisons between them (a source is either correct or incorrect on the object) and wish to determine a ranking among each kind (sources by trustworthiness, and objects by their difficulty). These two rankings should express a kind of mutual dependence – reminiscent of the principle of truth discovery described above – wherein, for example, sources correct on difficult objects should rank highly. A novel aspect of this particular problem is that there are no *direct* comparisons between entities of the same kind: sources do not go head-to-head, but must be ranked based on *indirect* patterns of correctness across objects.

This problem is an instance of a *tournament*. Tournaments consists of a set of *players* together with a *beating relation* between them, and have been widely studied in social choice theory. Obvious application domains include sports, but tournaments can be applied more widely (e.g. in voting, where candidate c "beats" d if a majority of voters prefer c to d). Many tournament solutions – producing either a set of winners or a full ranking of the players – have been proposed in the literature and studied axiomatically (González-Díaz, Hendrickx, and Lohmann 2014; Brandt, Brill, et al. 2016).

While our problem can be seen as a tournament, the bipartite nature and indirectness of comparisons between players to be ranked gives it a unique character worthy of dedicated study. We focus on a particular class of intuitive bipartite tournament ranking methods based on *chain editing*; a graph

[3] A more detailed overview of the literature will be given in Chapter 2.

modification problem mainly studied for its computational complexity properties (Yannakakis 1981) and recently suggested in the context of ranking by Jiao, Ravi, and Gatterbauer (2017). While their work focussed on algorithms and the complexity of variants of chain editing, we take an axiomatic approach in the social choice tradition in order to understand its properties as a ranking mechanism.

1.2 Logic-based Perspectives

Modal logic. In the logic-based part of the thesis, we start with a *modal logic* framework for reasoning about expertise. A modal language augments propositional logic with one or more *modalities*, which qualify the truth value of a proposition (Garson 2021). A typical interpretation is *necessity*: $\Box\varphi$ means the proposition φ is *necessarily true*, as opposed to merely being true. The dual notion of *possibility*, denoted $\Diamond\varphi$, is defined in terms of necessity by $\Diamond\varphi \equiv \neg\Box\neg\varphi$. That is, φ is *possibly true* if it is not necessarily false.

Various senses of "necessity" give rise to a rich landscape of logical systems, useful for modelling various domains. For example, in temporal logics $\Box\varphi$ means φ will necessarily always hold in the future (Goranko and Rumberg 2022). In deontic logics, $\Box\varphi$ means φ is a moral necessity; it is obligatory for φ to hold (McNamara and Van De Putte 2022). In epistemic and doxastic logics, $\Box\varphi$ means φ is necessary from the point of view of an agent's knowledge or beliefs about the world; one typically writes $\mathsf{K}\varphi$ or $\mathsf{B}\varphi$ instead of $\Box\varphi$ in these cases, to express that the agent "knows" or "believes" φ (Rendsvig and Symons 2021).

The most prominent semantic interpretation of modal formulas is based on *relational models* (also known as *Kripke models*). The key ingredients here are a set of *states* and a binary relation called the *accessibility relation*. The modal formula $\Box\varphi$ holds at a state x if φ holds at all states y accessible from x.[4] In this way the accessibility relation directly reflects which states are "possible" from others. Later, *neighbourhood semantics* were developed by Scott (1970) and Montague (1970), which generalise relational semantics. The idea is to replace the accessibility relation with a so-called *neighbourhood function*, which assigns to each state a collection of sets of states (called its *neighbourhood*). This neighbourhood explicitly lists the "necessary" propositions at each state: $\Box\varphi$ holds at x if the set of states where φ holds is a member of the neighbourhood of x (Pacuit 2017). Further still, *topological semantics* (also called the *interior semantics*) were first studied mathematically by McKinsey (1941) and McKinsey and Tarski (1944), and later reinterpreted in epistemic terms (see Özgün (2017, Chapter 1) for a historical overview). Here one equips the set of states with a topology, and $\Box\varphi$ holds at all points in the *interior* of the set where φ holds. That is, φ is necessary at a state x when there is an open neighbourhood of x in which φ holds at all points.

[4] A formal definition will be given in Section 4.3; for now we only wish to sketch the

Relational, neighbourhood and topological semantics have each proven to be useful for modelling notions related to our topic, such as information, trust, belief, and evidence.

On trust, Liau (2003) considered modalities $B_i\varphi$ (agent i believes φ), $I_{ij}\varphi$ (i acquires information φ from j) and $T_{ij}\varphi$ (i trusts j on φ), where trust has a neighbourhood interpretation. Dastani et al. (2004) extended this framework to consider how trust may be inferred, using notions of topics and *questions*. Herzig et al. (2010) introduced notions of trust and *reputation*, in a framework where trust is not primitive but built from beliefs, goals and actions. Sakama, Caminada, and Herzig (2014) studied beliefs, communication and intentions; this allowed them to consider different kinds of dishonesty, such as lying, half-truth and bullshitting. Further logical developments of trust were set out by Rodenhäuser (2014) and Tagliaferri (2019); see the references therein for a more thorough review of the literature on trust.

Interactions between knowledge, belief and *evidence* have been studied in epistemic logic. Moss and Parikh (1992) introduced the so-called subset space semantics to model knowledge and *epistemic effort*, which represents a kind of evidence-gathering performed by an epistemic agent, and has topological roots. Özgün (2017) further developed notions of evidence in epistemic logic from a topological perspective. In a series of papers, van Benthem and Pacuit (2011), van Benthem, Fernández-Duque, and Pacuit (2012), and van Benthem, Fernández-Duque, and Pacuit (2014) made extensive use of neighbourhood structures to model evidence, and in particular how inconsistent evidence can be combined to form beliefs.

The final strand of the modal literature we mention is *dynamic epistemic logic* (van Ditmarsch, van der Hoek, and Kooi 2008; Baltag and Renne 2016). Here modal operators describe *actions*: one has formulas of the form $[a]\varphi$ to express that φ is true after the action a is performed (Baltag and Renne 2016). Examples include public announcements (Plaza 2007), testimony (Holliday 2009) – both particularly interesting in the case of multiple agents – belief revision (Baltag and Smets 2008) and learning (Gierasimczuk 2009; Gierasimczuk 2010).

Our work in Chapter 4 combines elements from each of the above-surveyed areas of the literature to model *expertise* and its relation to relation to *truthfulness of information*. Specifically, we introduce a logical system with two new modalities: $\mathsf{E}\varphi$, meaning the source in question is an expert on φ, and $\mathsf{S}\varphi$, meaning the information φ is "sound" for the source to report. Informally, the latter notion means φ is true *up to lack of expertise*, i.e. the report becomes true when discarding parts of the statement on which the reporting source lacks expertise.

For example, suppose an economist reports that energy policies proposed by the government will stimulate economic growth and help tackle climate change. Since this goes beyond the expertise of the economist (who we assume

underlying ideas.

is not a climate expert), we should only take their comments on the economy into account. Our notion of soundness models this kind of filtering: whereas the statement in its entirety may be false (e.g. if the proposed policies are not in fact climate-friendly), the report is *sound* whenever the economist is correct on its economic content.

On the technical side, we use (a special case of) neighbourhood semantics for expertise, and topological semantics for soundness. We show in detail how such notions relate to knowledge in epistemic logic under relational semantics. Dynamic operators are also considered; we define an analogue of public announcements (called "sound announcements") and consider "expertise increase", which models the effects of a source increasing their domain of expertise by learning.

We also obtain axiomatisation results for various classes of expertise models. Note that "axiom" here has a different meaning to the axioms of social choice theory. In a logical system, axioms are formulas which – together with rules of inference – give rise to a notion of *syntactic entailment* or *proof*: one writes $\Gamma \vdash \varphi$ if φ can be derived using the axioms and inference rules from the assumptions in Γ. Axioms therefore form the building blocks of (syntactic) reasoning in a logic. There is also the dual notion of *semantic entailment*: $\Gamma \models \varphi$ if for every model and every state x, if all formulas in Γ hold at x then φ also holds at x.

The task of choosing axioms such that these two notions of entailment coincide is precisely what it means to find an axiomatisation. The implication $\Gamma \vdash \varphi \implies \Gamma \models \varphi$ is called *soundness*, and says that anything one can prove syntactically is actually true according to the semantics. The converse implication $\Gamma \models \varphi \implies \Gamma \vdash \varphi$ is called (strong) *completeness*;[5] this direction is typically more difficult to prove, and says that anything which is true semantically can indeed be proved.

By restricting to a sub-class of models, one obtains a different (and stronger) notion of semantic entailment. For example, one has $\Box\varphi \models \Box\Box\varphi$ when restricting to the class of *transitive* relational models, but not in general. Generally speaking, restricting the class of models under consideration requires more axioms to ensure completeness. In our case, we will consider several classes of expertise models in which various assumptions are placed on the nature of the source's expertise (for instance, the ways in which they may combine their expertise on separate pieces of information), and introduce additional axioms in each case.

Belief change. Whereas Chapter 4 proceeds in the tradition of modal epistemic logic, Chapter 5 takes inspiration from the literature on *belief change* (Booth and Meyer 2011; Hansson 2022; Fermé and Hansson 2018). This research area concerns how a rational agent should change its beliefs – represented by a logically closed set of formulas K[6] – in response to some operation. In belief

[5] *Weak* completeness is the special case of strong completeness in which $\Gamma = \emptyset$.

contraction, the agent must remove some formula α from its belief set, obtaining new beliefs $K \dot{-} \alpha$. In belief *revision*, the agent adjusts their beliefs to incorporate α, with the revised belief set denoted by $K * \alpha$. Other operations include *update* (Katsuno and Mendelzon 1992), where the agent must change their beliefs to account for changes in the outside world, and *merging*, where inputs from several sources – not necessarily consistent with one another – must be combined (Konieczny and Pino Pérez 2002).

Note that unlike in epistemic and doxastic logics, where belief is represented at the "object-level" via formulas $\mathsf{B}\varphi$, belief change research typically represents beliefs at the "meta-level" via sets of formulas (typically propositional formulas).

A common principle guiding belief change methods is *minimal change*: an agent's belief after contraction/revision should remain as close to the initial beliefs as possible. However, such minimal change may often be carried out in several ways. For example, consider an agent who believes p and $p \to q$. Assuming beliefs are closed under logical consequence, the agent also believes q. If the agent now learns q is false, i.e. has to revise by $\neg q$, there are two options: drop the belief in p and retain $p \to q$, or maintain p but drop $p \to q$. From a purely logical point of view, both are viable revision strategies. With this in mind, one cannot single out a single contraction/revision operator. Instead, the dominant approach is to set out *rationality postulates* which constrain the operation in question (much like the axioms of social choice theory, described in Section 1.1). Additional structure (e.g. preferences over formulas or propositional worlds) is then used to define specific operators within the bounds of the postulates.

The seminal work of Alchourrón, Gärdenfors, and Makinson (1985) set out postulates for contraction and revision. The influence of this work is such that belief revision in this framework is now called *AGM revision*, and the postulates the *AGM postulates*, named after its originators. For a fixed belief set K, the postulates are as follows.[7]

Closure. $K = \mathrm{Cn}(K)$

Success. $\alpha \in K * \alpha$

Inclusion. $K * \alpha \subseteq \mathrm{Cn}(K \cup \{\alpha\})$

Vacuity. If $K \cup \{\alpha\}$ is consistent, then $K * \alpha = \mathrm{Cn}(K \cup \{\alpha\})$

Consistency. If α is consistent, then $K * \alpha$ is consistent

[6]But note that this is not the only way to represent beliefs. For iterated change, abstract "epistemic states" are used instead of belief sets (Darwiche and Pearl 1997). In belief *base* change (Hansson 1999b), "belief bases" are again sets of formulas but are not necessarily closed under logical consequence.

[7]Note that some authors refer to the first six postulates as the "basic postulates" postulates, and the final two as the "supplementary postulates". In this thesis, "AGM postulates" refers to all eight postulates.

Extensionality. If $\alpha \equiv \beta$, then $K * \alpha = K * \beta$

Subexpansion. $K * (\alpha \wedge \beta) \subseteq \mathrm{Cn}\left((K * \alpha) \cup \{\beta\}\right)$

Superexpansion. If $(K*\alpha) \cup \{\beta\}$ is consistent, then $K*(\alpha \wedge \beta) \supseteq \mathrm{Cn}\left((K * \alpha) \cup \{\beta\}\right)$

An operator $*$ satisfying the postulates is called an *AGM operator* for K. Here $\mathrm{Cn}(\cdot)$ denotes logical consequence. Note that **Vacuity** goes some way to formalising the idea of minimal change: if the α is already consistent with current beliefs, the agent should just add α and close under logical consequence.

In our context, an important postulate which underlies the assumptions of the AGM framework is **Success**. As its name suggests, this says the revision process was successful: α is indeed believed after revision by α. Consequently, the framework assumes the information by which to revise is *completely reliable*. While clearly useful in some contexts, this severely limits the application scenarios of AGM revision. *Non-prioritised* revision subsequently arose to model situations where the incoming information α is not prioritised over existing beliefs K. Various approaches were surveyed by Hansson (1999a). Typically some extra structure accompanies a revision operator in order to decide to what extent the new information is accepted. For instance, *screened revision* (Makinson 1997) maintains a subset $A \subseteq K$ of "core" beliefs which are untouchable; α is only accepted if it is consistent with core beliefs. A similar construction for iterated revision was proposed by Booth (2005). *Selective revision* (Fermé and Hansson 1999) adds a pre-processing step to AGM revision: a so-called "selection function" f is used to weaken the incoming information α, and the revised belief set is $K*f(\alpha)$, where $*$ is a (prioritised) AGM revision operator. In *credibility-limited revision* (Hansson et al. 2001) one considers a set $\mathcal{C}$ of "credible" formulas; on receiving some $\alpha \in \mathcal{C}$ one revises by α with an AGM operator as usual, but if $\alpha \notin \mathcal{C}$ beliefs are unchanged.

Note that screened revision uses extra information about the *agent*, whereas selective and credibility-limited revision use extra information about the *information source*. Indeed, the selection function f can be used to filter out parts of the input α on which the source is deemed to be trustworthy. Booth and Hunter (2018) take this idea further, introducing a specialisation of selective revision in which the selection function is derived from an explicit representation of trust in the source. The idea is similar for credibility-limited revision, where $\mathcal{C}$ represents the formulas on which the source is trusted.

However, these approaches share a common deficiency: the non-prioritisation mechanism remains fixed. That is, the trustworthiness of the reporting source itself it not subject to revision. This is problematic in scenarios where information sources are not well-known up front, and becomes especially important when dealing with multiple sources. For example, consider conflicting reports on a breaking news story posted on Twitter by unfamiliar users X and Y. Initially we may be reluctant to commit to believing either, not knowing much about their expertise. As time goes on, however, more information becomes available. Perhaps a consensus emerges around the report of Y, or a report from a known, trusted source validates Y. In this case it may be natural to revise beliefs not only on the news in question, but on the *expertise of* X *and* Y *themselves*.

We take steps to resolve the situation by using the framework for expertise of Chapter 4 as the logical background for a belief change problem with non-expert sources. Our operators take a sequence of reports from multiple sources, and output a belief set in the extended language with expertise and soundness modalities. In keeping with the AGM paradigm, beliefs and trustworthiness are expressed on the meta-level. By using the extended language of expertise we unify the trust and belief aspects in a common logical language.[8] Beliefs about the world are expressed as propositional formulas as usual, and trust is expressed via *belief in expertise*. That is, a belief change agent trusts a source X on a formula φ if $\mathsf{E}_{\mathsf{X}}\varphi$ is included in its belief set.

Our approach is mainly postulational: we put forward a collection of postulates for expertise-and-belief revision and offer a number of constructions satisfying the postulates. Crucially we do *not* require **Success**, and our operators cope with inconsistent reports. Formally, the framework is closer to belief merging à la Konieczny and Pino Pérez (2002) than AGM-style revision, and a detailed comparison with merging is given in Section 5.6.

Learning. While AGM revision tells us how to revise beliefs in a rational and minimal way, it says nothing about whether the revised beliefs become *closer to the truth*. Indeed, belief revision theory does not include a model of the "true" state of the world, and thus the question of verisimilitude cannot be addressed without extending the framework in some way.

On an intuitive level, the conservative principle of minimal change – as expressed by the AGM postulates – may even conflict with pursuit of the truth: sometimes radical changes are required (Kelly, Schulte, and Hendricks 1997). For example, consider a conservative agent who strongly believes a biased coin is in fact fair. Then a sequence of 1000 consecutive heads, while vanishingly unlikely, is nevertheless *consistent* with a fair coin, and our agent will see no reason to change beliefs.

Fortunately, not *all* AGM operators embody conservatism to this extreme extent. Kelly, Schulte, and Hendricks (1997) analysed AGM revision from the point of view of formal learning theory (Jain et al. 1999; Gierasimczuk 2010, §2.1), and showed that AGM operators are *universal*, in the sense that if the truth can be found by any learning method at all, it can be found by an AGM operator. However, the initial beliefs need to be chosen carefully to avoid situations like the one described above. Specific AGM operators from the literature were also studied for their learning-theoretic properties (Kelly 1998; Kelly 1999). In a similar framework, Gierasimczuk (2010), Baltag, Gierasim-

[8]See the work of Yasser and Ismail (2020) – which is discussed in detail in Section 5.6 – for an alternative approach in which trust and belief are treated separately.

czuk, and Smets (2019), and Baltag, Gierasimczuk, and Smets (2016) studied various belief revision methods in relation to learning, and again proved universality of AGM learning.

While our belief change operators in Chapter 5 are inspired by the AGM paradigm, there is a crucial difference: we expressly aim to handle non-expert – and therefore *false* – information. In Chapter 6 we adapt the truth-tracking idea of Baltag, Gierasimczuk, and Smets (2019) to this new setting, and formalise the notion of *learning* from non-expert sources. We will investigate the extent to which learning is actually possible, both in terms of learning the true facts of the world and learning the true extent of the expertise of the sources themselves. In this respect the learning problem is similar to truth discovery, as studied in Chapter 2. A universality result of sorts is also available in our setting, and we offer a precise postulational characterisation of such truth-tracking methods. It turns out the some (but not all) of our example operators from the belief change problem in Chapter 5 are truth-tracking; this shows that the rationality postulates *are* compatible with learning from non-experts.

1.3 Contributions

The novel contributions of this thesis may be broken down per-chapter as follows.

Chapter 2.

- A formal framework for truth discovery suitable for axiomatic analysis in the style of social choice theory.
- The introduction of several axioms in this framework, the first impossibility theorems for truth discovery, and axiomatic analysis of a well-known truth discovery method from the literature.

Chapter 3.

- The definition of a new class of *bipartite tournaments* and the associated ranking problem.
- An in-depth study of *chain editing* as a ranking method for bipartite tournaments, including axiomatic analysis in the style of social choice theory.

Chapter 4.

- A modal logic framework to reason about the expertise of an information source, and "soundness" of information.

- Results establishing the connection between expertise and epistemic logic, so that expertise can be interpreted in terms of *knowledge*.
- Technical results on the mathematical properties of the logic of expertise, including axiomatisation results for several classes of expertise models.

Chapter 5.

- The statement of a new logic-based belief change problem for handling reports from multiple non-expert sources, in which we aim to determine what to believe both about the world and about the expertise of the sources themselves.
- New postulates and operators for this problem.

Chapter 6.

- A framework for truth-tracking with multiple non-expert sources, adapted from previous work from formal learning theory and belief revision.
- Mathematical analysis of when truth-tracking is possible with non-experts, and the extent to which one can learn the truth.

— 2 —
Truth Discovery

There is an increasing amount of data available in today's world, particularly from the web, social media platforms and crowdsourcing systems. The openness of such platforms makes it simple for a wide range of users to share information quickly and easily, potentially reaching a wide international audience. It is inevitable that amongst this abundance of data there are *conflicts*, where data sources disagree on the truth regarding a particular object or entity. For example, low-quality sources may mistakenly provide erroneous data for topics on which they lack expertise, or malicious sources may try to deliberately deceive.

Resolving such conflicts and determining the true facts is therefore an important task. Truth discovery has emerged as a set of techniques to achieve this by considering the *trustworthiness* of sources (Y. Li, Gao, et al. 2016; Gupta and Han 2011; Berti-Equille and Borge-Holthoefer 2015). The general principle is that true claims are those reported by trustworthy sources, and trustworthy sources are those that report believable claims. Note that there is a *mutual dependence* between the trust and belief parts of the problem, whereby highly trusted sources bestow credibility on their claims and vice versa. Application areas include real-time traffic navigation (Du et al. 2019), drug side-effect discovery (Ma, Meng, et al. 2017), and crowdsourcing and social sensing (D. Y. Zhang et al. 2016; D. Wang et al. 2012; Ma, Y. Li, et al. 2015).

For a simple example of a situation where trust can play an important role in conflict resolution, consider the following example.

Example 2.0.1. *Let o_1 and o_2 represent images for which crowdsourcing workers are asked to label $+$ or $-$ (in the truth discovery terminology, o_1 and o_2 are called* objects*; $+$ and $-$ are* values*). An object paired with a value is called a* claim*. Consider workers (the data sources) s, t, u and v who make claims reports as shown in Fig. 2.1. Without considering trust information, the label for o_1 appears a tie, with both options $+$ and $-$ receiving one vote from sources s and t respectively.*

Taking a trust-aware approach, however, we can look beyond object o_1 to consider the trustworthiness *of s and t. Indeed, when it comes to o_2, t agrees*

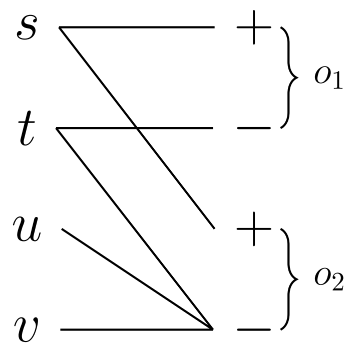


Figure 2.1: Illustrative example of a truth discovery problem, with sources s, t, u, v and objects o_1, o_2, each with associated values $+$ and $-$.

with two extra sources u and v, whereas s disagrees with everyone. In principle there could be many extra sources here instead of just two, in which case the effect would be even more striking. We may therefore postulate that s is less trustworthy *than t. Returning to o_1, we see that the label $-$ is supported by a more trustworthy source, and conclude that it should be accepted over $-$.*

Many truth discovery algorithms have been proposed in the literature with a wide range of techniques used, e.g. iterative heuristic-based methods (Pasternack and Roth 2010; Galland et al. 2010), probabilistic models (Yin, Han, and Yu 2008), maximum likelihood estimation and optimisation-based methods (Y. Li, Q. Li, et al. 2016), and neural network models (Kotonya and Toni 2020; Marshall, Argueta, and D. Wang 2017; Y. Wang et al. 2018). It is common for such algorithms to be evaluated empirically by running them against real-world or synthetic datasets for which the true facts are already known; this allows *accuracy* and other metrics to be calculated, and permits comparison between algorithms (see (Waguih and Berti-Equille 2014) for a systematic empirical evaluation of this kind). This may be accompanied by some theoretical analysis, such as calculating run-time complexity (Gupta and Han 2011), proving convergence of an iterative algorithm (Yin and Tan 2011), or proving convergence to the "true" facts under certain assumptions on the distribution of source trustworthiness (Xiao, Gao, et al. 2016; Xiao 2018; Ghosh, Kale, and McAfee 2011).

A limitation of this kind of analysis is that the results only apply narrowly to particular algorithms, due to the assumptions made (for instance, that claims from sources follow a particular probability distribution). Such assumptions can be problematic in domains where the desired truth is somewhat "fuzzy"; for example, image classification problems and determining the copyright status of books.[1]

In this work we take first steps towards a more general approach, in which we aim to study truth discovery without reference to any specific methodology

[1] https://www.nytimes.com/2019/08/19/technology/amazon-orwell-1984.html

or probabilistic framework. To do so we note the similarities between truth discovery and problems such as judgement aggregation (Endriss 2016), voting theory (Zwicker 2016) and ranking and recommendation systems (Altman and Tennenholtz 2008; Altman and Tennenholtz 2005; Andersen et al. 2008; Tennenholtz 2004) in which the *axiomatic approach* of social choice has been successfully applied. In taking the axiomatic approach one aims to formulate *axioms* that encode intuitively desirable properties that an algorithm may possess. The interaction between these axioms can then be studied; typical results include *impossibility results*, where it is shown that a set of axioms cannot hold simultaneously, and *characterisation results*, where it is shown that a set of axioms are uniquely satisfied by a particular algorithm.

Such analysis brings a new *normative* perspective to the truth discovery literature. This complements empirical evaluation: in addition to seeing how well an algorithm performs in practise on test datasets, one can check how well it does against theoretical properties that any "reasonable" algorithm should satisfy. The satisfaction (or failure) of such properties then shines new light on the *intuitive behaviour* of an algorithm, and may guide development of new ones.

With this in mind, we develop a framework for truth discovery in which axioms can be formulated, and go on to give impossibility results and an axiomatic characterisation of a baseline voting algorithm. We also define the class of *recursive* truth discovery algorithms, which includes most examples from the literature. We outline several specific examples: *Sums* (Pasternack and Roth 2010), *TruthFinder* (Yin, Han, and Yu 2008) and *CRH* (Y. Li, Q. Li, et al. 2016), and analyse Sums in more detail with respect to the axioms. Surprisingly, Sums fails some crucial axioms, which leads us to introducing a modified version with better axiomatic properties.

However, as a first step towards a social choice perspective of truth discovery, our framework involves a number of simplifying assumptions not commonly made in the truth discovery literature.

- **Collusion.** Our axioms assume sources act independently, in that there is no *collusion* or *copying* (Dong, Berti-Equille, and Srivastava 2009) among sources.

- **Object correlations.** We do not model *correlations* between the objects of interest in the truth discovery problem (Yang, Bai, and Liu 2019). For example, the crowdsourcing example it may be known in advance that objects o_1 and o_2 are similar, so that their true labels are correlated; this cannot be expressed in our framework.

- **Ordinal outputs.** For the most part, the outputs of our truth discovery methods consist of *rankings* of the sources and facts. Thus, we describe when a source is considered *more trustworthy* than another, but do not assign precise numerical values representing trustworthiness. This breaks with tradition in the truth discovery literature, but is a common point of view in social choice theory.

While this is something of a simplification compared to the current body of work in truth discovery, we argue that the problem is non-trivial even in this simplified setting, and that interesting axioms can still be put forth. The framework as set out here lays the groundwork for these assumptions to be lifted in future work.

Contributions. The primary contribution of this chapter is a mathematical framework for truth discovery, which allows for axiomatic analysis of truth discovery algorithms in the style of social choice theory. We introduce several axioms – many inspired by similar axioms in the social choice literature – which to date have not been considered in relation to truth discovery. This leads to the first impossibility theorems for truth discovery. Moreover, we observe that one particularly well-known method fails one of our axioms, and propose a modification to resolve this issue.

This chapter is a significantly re-worked version of previously published work (Singleton and Booth 2020; Singleton and Booth 2022a).

2.1 Preliminaries

In this section we give the basic definitions which form our formal framework.

Input. Intuitively, a truth discovery problem consists of a number of *sources* and a number of *objects* of interest. Each source provides a number of *claims*, where a claim is comprised of an object and a *value*. Different sources may give conflicting claims by providing different values for the same object. For simplicity, we only consider categorical values in this work. Note that while this restriction is made in some approaches in the literature (Pasternack and Roth 2010; Yin, Han, and Yu 2008; D. Wang et al. 2012; Dong, Berti-Equille, and Srivastava 2009; L. Zhang et al. 2018), in general truth discovery methods also handle continuous values (Y. Li, Q. Li, et al. 2016; Xiao, Gao, et al. 2016).

To formalise this, let $\mathbb{S}$, $\mathbb{O}$ and $\mathbb{V}$ be infinite, disjoint sets, representing the possible sources, objects and values. The input to the truth discovery problem is a *network*, defined as follows.

Definition 2.1.1. *A* truth discovery network *is a tuple* $N = (S, O, D, R)$*, where*

- $S \subseteq \mathbb{S}$ *is a finite set of* sources*.*
- $O \subseteq \mathbb{O}$ *is a finite set of* objects*.*
- $D = \{D_o\}_{o \in O}$ *are the* domains *of the objects, where each* $D_o \subseteq \mathbb{V}$ *is a finite set of values. We write* $V = \bigcup_{o \in O} D_o$*.*

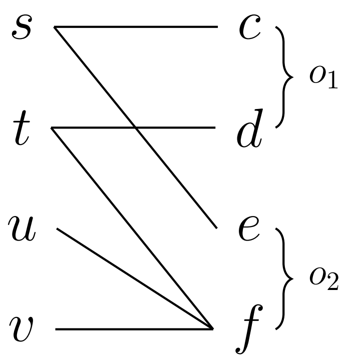


Figure 2.2: Claim-centric presentation of the network described in Fig. 2.1 and Example 2.0.1.

- *$R \subseteq S \times O \times V$ is a set of* reports.

such that

1. *For each $(s, o, v) \in R$, we have $v \in D_o$.*
2. *If $(s, o, v) \in R$ and $(s, o, v') \in R$, then $v = v'$.*

Note that while $\mathbb{S}$, $\mathbb{O}$ and $\mathbb{V}$ are infinite, each network is finite. The set R is the core data associated with the network: we interpret $(s, o, v) \in R$ as source s claiming that v is the true value for object o. Constraint (1) says that all claimed values are in the domain of the relevant object. Constraint (2) is a basic consistency requirement: a source cannot provide distinct values for a single object. That is, a source provides *at most one value* per object. Thus, while sources may be in conflict with *other sources*, they are not in conflict with themselves. While this is a simplifying assumption, we argue the truth discovery problem is still rich enough when conflicts only arise between distinct sources.

When a network N is understood, we often write S, O, D and R to implicitly refer to the components of N. Any decoration applied to N will also be applied to its components (e.g. N' has sources S', $\hat{N}$ has sources $\hat{S}$ etc...). If necessary, we write S_N, O_N, D_N and R_N to make the dependence on N explicit.

A *claim* is a pair $c = (o, v)$, where $o \in O$ and $v \in D_o$. We write $\text{obj}(c) = o$ in this case, and let C denote the set of all possible claims in a network N, i.e.

$$C = \{(o, v) \mid o \in O, v \in D_o\}.$$

Note that not every claim is necessarily reported by some source. With slight abuse of notation, we write (s, c) for the report (s, o, v). Then R can be viewed as a subset of $S \times C$, i.e. a relation between sources and claims. In fact, we will take this claim-centric view in the remainder of the chapter, with objects and values only playing a role insofar as they tell us which claims are in conflict with one another.

Example 2.1.1. *The network illustrated in Fig. 2.1 and Example 2.0.1 is given by* $S = \{s, t, u, v\}$, $O = \{o_1, o_2\}$ *and* $D_{o_1} = D_{o_2} = \{+, -\}$. *Labelling the claims* $c = (o_1, +)$, $d = (o_1, -)$, $e = (o_2, +)$ *and* $f = (o_2, -)$, *we have* $R = \{(s, c), (s, e), (t, d), (t, f), (u, f), (v, f)\}$. *This "claim-centric" view of the network is shown in Fig. 2.2, where the values + and − are suppressed.*

Example 2.1.1 highlights a special case of our framework: the "binary" case in which the domain of each object consists of two values $D_o = \{+, -\}$. In this case we can think of each object as a propositional variable. This brings us close to the setting studied in *judgement aggregation* (Endriss 2016) and, specifically (since sources do not necessarily provide a claim for each object) to the setting of *binary aggregation with abstentions* (Christoff and Grossi 2017; Dokow and Holzman 2010). An important difference, however, is that for simplicity we do not assume any *constraints* on the possible configurations of true claims across objects. That is, any combination of truth values is feasible. In judgement aggregation such an assumption has the effect of neutralising the impossibility results that arise in that domain (see e.g., (Christoff and Grossi 2017)). We shall see later that that is not the case in our setting.

Notation. We introduce some notation to extract information about a network. For $c \in C$ and $s \in S$, write

$$\begin{aligned} \mathrm{src}_N(c) &= \{s \in S \mid (s, c) \in R\}, \\ \mathrm{cl}_N(s) &= \{c \in C \mid (s, c) \in R\}. \end{aligned}$$

The set of sources making a claim on object o is

$$\mathrm{src}_N(o) = \bigcup \{\mathrm{src}_N(c) \mid c \in C, \mathrm{obj}(c) = o\}.$$

The claims associated with o are

$$\mathrm{cl}_N(o) = \{c \in C \mid \mathrm{obj}(c) = o\}.$$

The set of claims in conflict with a given claim $c = (o, v)$, i.e. claims for o with a value other than v, is denoted by

$$\mathrm{conflict}_N(c) = \{(o, v') \mid v' \in D_o \setminus \{v\}\}.$$

The "antisources" of c are then defined to be the sources for claims conflicting with c:

$$\mathrm{antisrc}_N(c) = \bigcup \{\mathrm{src}_N(d) \mid d \in \mathrm{conflict}_N(c)\}.$$

Note that property (2) in the definition of a network ensures $\mathrm{src}_N(c) \cap \mathrm{antisrc}_N(c) = \emptyset$.

Output. With the input defined, we now come to the output of the truth discovery problem. The primary goal is to produce an assessment of the trustworthiness of the sources, and the *true values* for the objects. Approaches differ regarding values: some truth discovery methods output only a single value for each object (Y. Li, Q. Li, et al. 2016; Ding, Gao, and Xu 2016; Yang, Bai, and Liu 2018), whereas others give an assessment of the believability (or confidence, probability etc…) of *each claim* (o, v) (Yin, Han, and Yu 2008; Pasternack and Roth 2010; Galland et al. 2010; Zhi et al. 2015; D. Y. Zhang et al. 2016; L. Zhang et al. 2018). We opt for the latter, more general, approach.

On the specific form of these assessments, we face a tension between the social choice and truth discovery perspectives. In social choice theory, one generally looks at *rankings*: e.g. the ranking of candidates in an election result according to a voting rule. Consequently, axioms are generally *ordinal properties*, which constrain how candidates (for example) compare *relative to each other*. In contrast, truth discovery methods universally use *numeric values*. This is more convenient for defining and using truth discovery methods in practise, and induces a ranking by simply comparing the numeric scores. The magnitude of the differences between scores also gives information about *confidence* in distinguishing sources and claims.

However, numeric scores are often not comparable between different methods (for example, some methods output probabilities, whereas others are interpreted as weights which may take negative values) and in general may not carry any semantic meaning at all. This means that meaningful axioms for truth discovery should not refer to specific numeric scores, but only the ranking they introduce.

We will ultimately take a hybrid approach: our methods and examples will be defined in terms of numeric scores, but the axioms will only refer to ordinal properties. This approach is summarised succinctly by Altman and Tennenholtz (2008), who write of ranking systems: “We feel that the numeric approach is more suitable for defining and executing ranking systems, while the global ordinal approach is more suitable for axiomatic classification.”

An *operator* maps each network to score and claim scores.

Definition 2.1.2. *A* truth discovery operator *T maps each network N to a function $T_N : S_N \cup C_N \to \mathbb{R}$.*

Intuitively, the higher the score $T_N(s)$ for a source $s \in S$, the *more trustworthy* s is, according to T on the basis of N. Similarly, the higher $T_N(c)$ for a claim $c \in C$, the *more believable* c is deemed to be. We define the source and claim rankings associated with T and N by

$$
\begin{aligned}
s \sqsubseteq_N^T s' &\iff T_N(s) \leq T_N(s'), \\
c \preceq_N^T c' &\iff T_N(c) \leq T_N(c').
\end{aligned}
$$

Then $s \sqsubseteq_N^T s'$ if s' is at least as trustworthy as s, and similar for $\preceq_N^T$. Note that $\sqsubseteq_N^T$ and $\preceq_N^T$ are total preorders.[2] We denote the strict parts by $\sqsubset_N^T$ and

$\prec_N^T$ respectively, and the symmetric parts by $\simeq_N^T$ and $\approx_N^T$. We omit the sub- and super-scripts when N and T are clear from context.

Given that our axioms will only refer to the rankings produced by operators, two operators yielding exactly the same rankings – possibly with different scores – appear the same with respect to axiomatic analysis. We say operators T and T' are *ranking equivalent*, denoted $T \sim T'$, if for all networks N we have $\sqsubseteq_N^T = \sqsubseteq_N^{T'}$ and $\preceq_N^T = \preceq_N^{T'}$.

In Section 2.2 we will introduce operators defined as the limit of an iterative procedure. To allow for possible non-convergence we also consider *partial operators*, which assign a mapping $T_N : S \cup C \to \mathbb{R}$ for only a subset of networks.

2.2 Example Operators

In this section we capture several example operators from the literature in our framework: a baseline *voting* method and its generalisation to *weighted* voting, *Sums* (Pasternack and Roth 2010), *TruthFinder* (Yin, Han, and Yu 2008) and *CRH* (Y. Li, Q. Li, et al. 2016). As is the case with many methods in the literature, the latter three methods operate iteratively: starting with an initial estimate, scores are repeatedly updated according to some procedure until convergence. Typically the update procedure is recursive, with source scores being updated on the basis of the current claims scores, and vice versa. To simplify the definition and analysis of such methods, we will introduce the class of *recursive operators*.

2.2.1 Voting

It is common in the literature to evaluate truth discovery methods against a non-trust-aware method, such as a simple voting procedure.[3] Here we consider each source to "vote" for their claims, and claims are ranked according to the number of votes received, i.e. by $|\operatorname{src}_N(c)|$. While this ignores the trust aspect of truth discovery entirely, this method will be useful for us as an axiomatic baseline. For example, axioms which aim to address the trust aspect should not hold for voting, and an axiom referring to the ranking of claims may be too strong if it does hold for voting.

Definition 2.2.1. T^{vote} *is the operator defined by*

$$\begin{aligned} T_N^{\text{vote}}(s) &= 1, \\ T_N^{\text{vote}}(c) &= |\operatorname{src}_N(c)|. \end{aligned}$$

[2] A total preorder is a transitive, reflexive and complete binary relation.

[3] This is often called *majority voting* in the truth discovery literature (e.g. (Y. Li, Gao, et al. 2016; Xiao and S. Wang 2015; Y. Li, Q. Li, et al. 2016)), but using the terminology of social choice theory it is better described as *plurality voting*.

Applying T^{vote} to the network in Fig. 2.2, we have that all sources rank equally ($s \simeq t \simeq u \simeq v$) and $c \approx d \approx e \prec f$.

The problem with T^{vote} is that all reports are equally weighted. If we have a mechanism by which sources can be weighted by trustworthiness, the idea behind voting may still have some merit. We define *weighted voting* as follows.

Definition 2.2.2. *A* weighting w *maps each network* N *to a function* $w_N : S \to \mathbb{R}$*. The associated* weighted voting *operator* T^w *is defined by*

$$\begin{aligned} T^w_N(s) &= w_N(s), \\ T^w_N(c) &= \sum_{s \in \operatorname{src}_N(c)} w_N(s). \end{aligned}$$

Note that T^{vote} arises via the weighting $w_N \equiv 1$. Note that a weighting is essentially just half of a truth discovery operator, where we only output scores for sources. This is completed to an operator T^w by letting the score for a claim be the sum of the weights of its sources. Note also that we allow the possibility of "untrustworthy" sources with $w_N(s) < 0$. Reports from such sources *decrease* the credibility of a claim.

Example 2.2.1. *Set*

$$w^{\text{agg}}_N(s) = \sum_{c \in \operatorname{cl}_N(s)} \frac{|\operatorname{src}_N(c)|}{|\operatorname{cl}_N(s)|}.$$

Then the weight assigned to a source s *is the average number of sources agreeing with the claims of* s*. We call the corresponding operator* Weighted Agreement*. Taking* N *from Fig. 2.2, we have* $w^{\text{agg}}_N(s) = 1$*,* $w^{\text{agg}}_N(t) = 2$*,* $w^{\text{agg}}_N(u) = 3$*,* $w^{\text{agg}}_N(v) = 3$*. Consequently,*

$$\begin{aligned} T^{w^{\text{agg}}}_N(c) &= w^{\text{agg}}_N(s) = 1, \\ T^{w^{\text{agg}}}_N(d) &= w^{\text{agg}}_N(t) = 2, \\ T^{w^{\text{agg}}}_N(e) &= w^{\text{agg}}_N(s) = 1, \\ T^{w^{\text{agg}}}_N(f) &= w^{\text{agg}}_N(t) + w^{\text{agg}}_N(u) + w^{\text{agg}}_N(v) = 8, \end{aligned}$$

yielding the rankings $s \sqsubset t \sqsubset u \simeq v$ *and* $c \approx e \prec d \prec f$*. Note that claim* d *fares better here than with* T^{vote} *due to its association with source* t*, who is more trustworthy than* s*.*

As we will see in Section 2.4, some operators do not correspond exactly to a weighting w, but give rise to the same rankings. Let us say an operator T is *weightable* if there exists a weighting w such that $T \sim T^w$. Given that weighted voting expresses a clear relationship between source and claim scores, this notion will simplify some aspects of axiomatic analysis later.

2.2.2 Recursive Operators

To capture the mutual dependence between trust in sources and belief in claims, truth discovery methods generally involve recursive computation (Pasternack and Roth 2010; Yin, Han, and Yu 2008; Yang, Bai, and Liu 2019; Du et al. 2019; L. Zhang et al. 2018; Y. Li, Q. Li, et al. 2016; Galland et al. 2010; Zhi et al. 2015). Claim scores are updated on the basis of currently estimated source scores, before claim scores are updated on the basis of the new sources scores. If this process converges, the limiting scores should be a fixed-point of the update procedure, reflecting the desired mutual dependence. To formalise this idea, we define recursive operators.

Definition 2.2.3. *A* recursive scheme *is a tuple* $(\mathcal{D}, T^0, U)$*, where*

- $\mathcal{D}$ *is a set of operators.*
- $T^0 \in \mathcal{D}$ *is the* initial operator*.*
- $U : \mathcal{D} \to \mathcal{D}$ *is the* update function*.*

A recursive scheme converges *to an operator* T^* *if for all networks* N *and all* $z \in S \cup C$*,* $\lim_{n\to\infty} U^n(T_0)_N(z) = T^*_N(z)$*. In this case* T^* *is said to be the* limit *of the scheme.*

The main component of interest here is the update function U, which describes how the scores of one iteration are transformed to obtain scores for the next. The domain of operators $\mathcal{D}$ is used for technical reasons; for example, some operators need to exclude the trivial operator in which scores are identically zero in order for U to be well-defined.

Note that the limit operator T^* is unique, when it exists. We can consider any scheme to converge to a *partial* operator T^*, defined on the networks N such that $\lim_{n\to\infty} U^n(T_0)_N(z)$ exists for all $z \in S \cup C$. We now consider examples of recursive operators from the literature.

Sums. Sums (Pasternack and Roth 2010) is a simple and well-known operator adapted from the *Hubs and Authorities* (Kleinberg 1999) algorithm for ranking web pages. The premise is to extend the linear sum of weighted voting to both claim and source scores: we update the score of each source as the sum of the scores of its claims, and update the score of each claim as the sum of the scores of its sources. To prevent scores from growing without bound, they are normalised at each iteration by dividing by the maximum score (for sources and claims separately).

Definition 2.2.4. Sums *is the recursive scheme* $(\mathcal{D}, T^0, U)$*, where* $\mathcal{D}$ *is the set of all operators with scores in* $[0,1]$*,* $T^0_N \equiv 1/2$*, and* $U(T) = T'$*, with*

$$T'_N(s) = \alpha \sum_{c \in \mathrm{cl}_N(s)} T_N(c),$$
$$T'_N(c) = \beta \sum_{s \in \mathrm{src}_N(c)} T'_N(s).$$

where $\alpha = 1/\max_{t\in S} \left|\sum_{c\in \mathrm{cl}_N(t)} T_N(c)\right|$ *and* $\beta = 1/\max_{d\in C} \left|\sum_{s\in \mathrm{src}_N(d)} T'_N(s)\right|$ *are normalisation factors (which we set to 0 if the denominator is 0). Write* T^{sums} *for the associated limit operator.*

Taking the network N from Fig. 2.2, one can show that $T^{\mathrm{sums}}_N(s) = 0$, $T^{\mathrm{sums}}_N(t) = 1$ and $T^{\mathrm{sums}}_N(u) = T^{\mathrm{sums}}_N(v) = \sqrt{2}/2 \approx 0.7071$, giving a source ranking $s \sqsubset u \simeq v \sqsubset t$. For claims, we have $T^{\mathrm{sums}}_N(c) = T^{\mathrm{sums}}_N(e) = 0$, $T^{\mathrm{sums}}_N(d) = \sqrt{2} - 1 \approx 0.4142$ and $T_N(f) = 1$, giving a claim ranking $c \approx e \prec d \prec f$. Note that the claim ranking is identical to that of Example 2.2.1. For sources, we see that t moves strictly upwards in the ranking compared to Example 2.2.1. Intuitively, this is because source t claims a superset of the claims of u and v, so receives more weight from its claims at each iteration.

TruthFinder. TruthFinder (Yin, Han, and Yu 2008) is a pseudo-probabilistic method, and was defined in the first work to introduce (and coin the phrase) truth discovery. It is formulated in a setting more general than ours: the authors suppose claims may *support* each other, as well as conflict, and that support of conflict may occur to varying degrees. Formally, each pair of claims c, c' has an "implication" value $\mathrm{imp}(c \to c') \in [-1, 1]$, where a negative value implies confidence in c should decrease confidence in c', and a positive value implies confidence in c should *increase* confidence in c'. In contrast, our framework assumes claims for the same object are mutually exclusive, so that all implications are negative. To express TruthFinder in our framework, we take $\mathrm{imp}(c \to c')$ to be $-\lambda$ if c and c' have the same object and 0 otherwise, for some fixed parameter $0 \leq \lambda \leq 1$.

Definition 2.2.5. *Given parameters* $\rho, \gamma \in (0,1)$ *and* $\lambda \in [0,1]$*,* TruthFinder *is the recursive scheme* $(\mathcal{D}, T^0, U)$*, where* $\mathcal{D}$ *is the set of operators with* $0 < T_N(s) < 1$ *for all* N *and* $s \in S$ *with* $\mathrm{cl}_N(s) \neq \emptyset$*,* $T^0 \equiv 0.9$*, and* $U(T) = T'$*, with*

$$T'_N(c) = \left[1 + \frac{\prod_{s\in \mathrm{src}_N(c)} (1 - T_N(s))^{\gamma}}{\prod_{t \in \mathrm{antisrc}_N(c)} (1 - T_N(t))^{\gamma\rho\lambda}}\right]^{-1}, \tag{2.1}$$

$$T'_N(s) = \sum_{c \in \mathrm{cl}_N(s)} \frac{T'_N(c)}{|\,\mathrm{cl}_N(s)|}. \tag{2.2}$$

We write T^{tf} *for the associated limit operator.*

We refer the reader to the original TruthFinder paper (Yin, Han, and Yu 2008) for the interpretation of ρ and γ. As described above, λ controls the amount to which conflicting claims play a role in the evaluation of a given claim. Of special interest is the case $\lambda = 0$, in which the denominator in (2.1) is 1. Note that in (2.1) we have unfolded the definitions of (Yin, Han, and Yu 2008) in order to obtain a single expression of $T'_N(c)$ in terms of the $T_N(s)$, at the expense of interpretability.

Let us return again to the network in Fig. 2.2. We take parameters $\rho = 0.5$ and $\gamma = 0.3$ (as per the experimental setup of Yin, Han, and Yu (2008)) and $\lambda = 0.5$. Assuming that TruthFinder does indeed converge on this network – as it appears to do empirically – we have $T^{\mathrm{tf}}_N(s) \approx 0.5067$, $T^{\mathrm{tf}}_n(t) \approx 0.6590$ and $T^{\mathrm{tf}}_N(u) = T^{\mathrm{tf}}_N(v) = 0.7510$, which gives the ranking $s \sqsubset t \sqsubset u \simeq v$ on the sources. We have $T^{\mathrm{tf}}_N(c) \approx 0.5328$, $T^{\mathrm{tf}}_N(d) \approx 0.5670$, $T^{\mathrm{tf}}_N(e) \approx 0.4807$ and $T^{\mathrm{tf}}_N(f) \approx 0.7510$, which gives the ranking $e \prec c \prec d \prec f$ on the claims. Note that the source ranking coincides with that of Example 2.2.1, and the claim ranking refines that of Example 2.2.1 and Sums by ranking e *strictly* worse than c. Intuitively, this occurs because e has more sources reporting the conflicting claim (namely, f) than c does. If we instead take $\lambda = 0$, so that sources for conflicting claims are not considered, then the ranking reverts to $c \approx e \prec d \prec f$ (and the source ranking remains the same).

CRH. Standing for "Conflict Resolution on Heterogeneous Data", CRH is an optimisation-based framework for truth discovery (Y. Li, Q. Li, et al. 2016). It is again set in a more general setting, in which a metric d_o is available to measure the distance between values in D_o, for each object o. The optimisation problem jointly chooses weights for each source and a value for each object, such that the weighted sum of d_o-distances from each source's claim on o is minimised.

To express CRH in our framework we use the "probabilistic" encoding of categorical variables as described in (Y. Li, Q. Li, et al. 2016, §2.4.1), where each categorical value is represented as a one-hot vector, and the source weight regularisation from (Y. Li, Q. Li, et al. 2016, Eq. (4)). We make a minor modification, however, by adding a small quantity ε to α_s and $T'_N(s)$ defined below; this ensures the logarithm in $T'_N(s)$ and the division in $T'_N(c)$ is well-defined and simplifies analysis of CRH later on.

Definition 2.2.6. *Given $\varepsilon > 0$,* CRH-ε *is the recursive scheme $(\mathcal{D}, T^0, U)$, where $\mathcal{D}$ is the set of operators with $0 \leq T_N(c) \leq 1$ for all N and $c \in C$,*

$$T^0_N(s) = 0, \qquad T^0_N(c) = \frac{|\operatorname{src}_N(c)|}{|S|}.$$

Table 2.1: Output rankings of the example operators on the network from Fig. 2.2.

Voting	$s \simeq t \simeq u \simeq v$	$c \approx d \approx e \prec f$
Weighted Agreement	$s \sqsubset t \sqsubset u \simeq v$	$c \approx e \prec d \prec f$
Sums	$s \sqsubset u \simeq v \sqsubset t$	$c \approx e \prec d \prec f$
TruthFinder	$s \sqsubset t \sqsubset u \simeq v$	$e \prec c \prec d \prec f$
TruthFinder ($\lambda = 0$)	$s \sqsubset t \sqsubset u \simeq v$	$c \approx e \prec d \prec f$
CRH-ε	$s \sqsubset t \sqsubset u \simeq v$	$c \approx e \prec d \prec f$

and $U(T) = T'$, where

$$T'_N(s) = \varepsilon - \log\left(\frac{\alpha_s}{\sum_{t \in S} \alpha_t}\right),$$
$$T'_N(c) = \frac{\sum_{s \in \mathrm{src}_N(c)} T'_N(s)}{\sum_{t \in S} T'_N(t)},$$

with

$$\alpha_s = \varepsilon + \sum_{c \in \mathrm{cl}_N(s)} \sum_{d \in \mathrm{cl}_N(\mathrm{obj}(c))} (T_N(d) - \mathbb{1}[d = c])^2.$$

The limit operator is denoted by $T^{\text{crh-}\varepsilon}$.[4]

Note that in the case where each source provides a report on *all* objects – which is the setting in which CRH was originally introduced – we have $\sum_{c \in \mathrm{cl}_N(o)} T'_N(c) = 1$. Consequently, T'_N gives rise to a probability distribution over claims for each object o. The term of the sum in α_s corresponding to c is the squared Euclidean distance between this distribution and the distribution put forward by source s, which places all the probability mass in their report c.

In the network from Fig. 2.2 with $\varepsilon = 10^{-5}$, we have $T_N^{\text{crh-}\varepsilon}(s) \approx 0.2577$, $T_N^{\text{crh-}\varepsilon}(t) \approx 1.4827$ and $T_N^{\text{crh-}\varepsilon}(u) = T_N^{\text{crh-}\varepsilon}(v) \approx 9.3567$, giving the source ranking $s \sqsubset t \sqsubset u \simeq v$. Note that this is the same ranking on sources as T^{tf} gives. For claims, we have $T_N^{\text{crh-}\varepsilon}(c) = T_N^{\text{crh-}\varepsilon}(e) \approx 0.0126$, $T_N^{\text{crh-}\varepsilon}(d) \approx 0.0725$ and $T_N^{\text{crh-}\varepsilon}(f) \approx 0.9874$, giving the ranking $c \approx e \prec d \prec f$; this is the same as T^{sums}.

Table 2.1 summaries the source and claim rankings for each example operator on the network N from Fig. 2.2.

2.3 The Axioms

Having laid out the formal framework, we now introduce axioms for truth discovery. Such axioms are formal properties an operator may satisfy, which encode intuitively desirable behaviour. Many of our axioms are adaptations of

[4] In the degenerate case $S = \emptyset$, we set $T_N \equiv 0$.

axioms for various problem in social choice theory (e.g. from voting (Zwicker 2016) and ranking systems (Altman and Tennenholtz 2008)), in which the axiomatic method has seen great success. We also consider standard social choice axioms which are *not* desirable for truth discovery, to highlight the differences with classical problems such as voting. We will later revisit the example operators of the previous section to see to what extent our axioms hold in practise.

2.3.1 Coherence

The guiding principle of truth discovery is that claims backed by trustworthy sources should be believed, and sources making believable claims are trustworthy. All truth discovery methods aim to implement this principle to some extent, and the examples of Section 2.2 illustrate several different approaches.

We aim to formulate this principle axiomatically as a *coherency* property relating the source ranking $\sqsubseteq$ and the claim ranking $\preceq$: sources making higher $\preceq$-ranked claims should rank highly in $\sqsubseteq$, and vice versa. To do so we adapt the idea behind the *Transitivity* axiom of Altman and Tennenholtz (2008) for ranking systems.

Now, a difficulty arises when considering how to compare the claims of two sources. For a simple example, suppose sources have either *low*, *medium* or *high* trustworthiness. How should we rank a claim c with one *medium* source versus a claim d with a *low* and a *high* source? In some situations we may want to prioritise the number of claims, so that d is preferred. In others we may want to avoid trusting *low* sources as much as possible, so that c is preferred. The third option of ranking c and d equally believable is also reasonable.

To avoid these ambiguous cases, we focus on scenarios where there is an "obvious" ordering between two sets of claims (or sources). For example, consider the network depicted in Fig. 2.3. Suppose an operator gives a source ranking $s \sqsubset u \sqsubset t \sqsubset v$. Note that claims c and d have the same number of sources. Moreover, we can pair up these sources one-to-one such that the source for c is less trustworthy than the corresponding source for d: we have $s \sqsubset u$ and $t \sqsubset v$. On aggregate, we may reasonably say that $\mathrm{src}_N(c)$ is less trustworthy (with respect to $\sqsubseteq$) than $\mathrm{src}_N(d)$. We should therefore have $c \prec d$; any operator violating this has failed to realise the dependence between source trustworthiness and claim believability. Similarly, this reasoning can be applied to the set of claims from two sources.

This will form the basis of our first set of axioms. First, we formalise the above idea of a one-to-one correspondence respecting a ranking.

Definition 2.3.1. *If $\leq$ is a relation on a set X and $A, B \subseteq X$, then A* precedes *B* pairwise *with respect to $\leq$ if*

$$\exists f : A \to B \text{ bijective s.t. } \forall x \in A : \; x \leq f(x). \tag{2.3}$$

Say A strictly precedes *B if A precedes B but B does not precede A.*

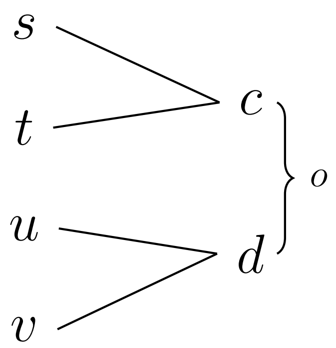


Figure 2.3: A network illustrating ***Claim-coherence****.*

If f satisfies the condition in (2.3), we say f *witnesses* the fact that A precedes B, and write $f : A \overset{\leq}{\Rightarrow} B$. Note that if $\leq$ is a preorder on X, the "precedes pairwise" relation is a preorder on 2^X. Indeed, it is reflexive (by considering the identity map $A \to A$, for each $A \subseteq X$) and transitive (if $f : A \overset{\leq}{\Rightarrow} B$ and $g : B \overset{\leq}{\Rightarrow} C$, then $g \circ f : A \overset{\leq}{\Rightarrow} C$). The strict pairwise order associated has a natural interpretation, as we now prove: there must exist some x in (2.3) for which the comparison is strict.

Proposition 2.3.1. *Suppose X is finite and $\leq$ is a total preorder on X. Then A strictly precedes B pairwise with respect to $\leq$ if and only if there is $f : A \overset{\leq}{\Rightarrow} B$ such that there is some $x_0 \in A$ with $x_0 < f(x_0)$.*

We need a preliminary lemma.

Lemma 2.3.1. *Suppose $\leq$ is a total preorder on a finite set X and $f : X \to X$ is an injective mapping such that $x \leq f(x)$ for all $x \in X$. Then $x \approx f(x)$ for all x.*

Proof. Take $x \in X$. Consider the sequence of iterates $(f^n(x))_{n \geq 1}$. Since this is an infinite sequence taking values in a finite set, there must be some point at which the sequence repeats, i.e. there are $n, k \geq 1$ such that $f^n(x) = f^{n+k}(x)$. Then $f(f^{n-1}(x)) = f(f^{n+k-1}(x))$, so injectivity gives $f^{n-1}(x) = f^{n+k-1}(x)$. Repeating this argument, we find $x = f^0(x) = f^k(x)$. By hypothesis, $f(x) \leq f^k(x)$, i.e. $f(x) \leq x$. Since $x \leq f(x)$ also, this gives $x \approx f(x)$ as required. □

Proof of Proposition 2.3.1. "if": Clearly A precedes B. Suppose for contradiction that this is not strict. Then there is some $g : B \overset{\leq}{\Rightarrow} A$. Note that $g \circ f$ is a bijection $A \to A$, and for all $x \in X$ we have $x \leq f(x) \leq g(f(x))$. By Lemma 2.3.1, $x \approx g(f(x))$. In particular, we have $f(x_0) \leq g(f(x_0)) \approx x_0$, but this contradicts $x_0 < f(x_0)$.

"only if": Suppose A strictly precedes B. Then there is some $f : A \overset{\leq}{\Rightarrow} B$. Note that f^{-1} is a bijection $B \to A$. Since B does not precede A, there must be some $y_0 \in B$ such that $y_0 \not\leq f^{-1}(y_0)$. By totality of $\leq$, we get $f^{-1}(y_0) < y_0$. Taking $x_0 = f^{-1}(y_0)$, we are done. □

We are now ready to state our first two axioms.

> **Claim-coherence**. If $\mathrm{src}_N(c)$ strictly precedes $\mathrm{src}_N(c')$ pairwise with respect to $\sqsubseteq^T_N$, then $c \prec^T_N c'$.
>
> **Source-coherence**. If $\mathrm{cl}_N(s)$ strictly precedes $\mathrm{cl}_N(s')$ pairwise with respect to $\preceq^T_N$, then $s \sqsubset^T_N s'$.

In words, **Claim-coherence** says that whenever we can pair up the sources for c and c' so that each source for c is less trustworthy than the corresponding source for c' (and *strictly* less, for at least one pair of sources), then c is strictly less believable than c'. Likewise, **Source-coherence** says that if the claims of s and s' can be paired up with the claims for s less believable than the claims for s', then s is strictly less trustworthy than s'.

Example 2.3.1. *Consider the network N from Fig. 2.2 again, and consider Sums. Recall that T^{sums} gives the source ranking $s \sqsubset u \simeq v \sqsubset t$, and claim ranking $c \approx e \prec d \prec f$.*

*Note that $\mathrm{src}_N(c) = \{s\}$ and $\mathrm{src}_N(d) = \{t\}$. Since $s \sqsubset t$, we have that $\{s\}$ strictly precedes $\{t\}$ with respect to $\sqsubseteq$. **Claim-coherence** therefore requires that $c \prec d$. Indeed, this does hold.*

*For **Source-coherence**, note that $\mathrm{cl}_N(s) = \{c, e\}$ and $\mathrm{cl}_N(t) = \{d, f\}$. Since $c \prec d$ and $e \prec f$, we see that $\mathrm{cl}_N(s)$ strictly precedes $\mathrm{cl}_N(t)$ with respect to $\preceq$. Accordingly, **Source-coherence** requires $s \sqsubset t$, which does hold.*

So, T^{sums} satisfies both coherence properties for this specific network. We will analyse T^{sums} and the other examples more generally in Section 2.4.

The reader may wonder why we only consider the *strict* pairwise relation in **Claim-coherence** (and **Source-coherence**). An alternative axiom might require that $c \preceq c'$ whenever $\mathrm{src}_N(s)$ precedes $\mathrm{src}_N(s')$ with respect to $\sqsubseteq$ (not necessarily strictly). However, this property implies that $c \approx c'$ whenever $\mathrm{src}_N(c) = \mathrm{src}_N(c')$. We have already seen an example operator where this does not hold: TruthFinder ranks $e \prec c$ in the network N from Fig. 2.2, but $\mathrm{src}_N(c) = \mathrm{src}_N(e) = \{s\}$. Intuitively, c and e are "tied" when it come to the quality of their own sources, but there are fewer sources *disagreeing* with c (the "antisources") than e. Stating our coherence properties in the strict form permits an operator to consider antisources in cases where there is no clear comparison on the basis of sources alone.

Having said this, an operator with **Claim-coherence** is limited in the extent to which it can take antisources into account. We formulate an antisource version of coherence in Section 2.3.5, and show that it is incompatible with **Claim-coherence** when taken with some other basic axioms.

2.3.2 Symmetry

A standard class of axioms in social choice theory express *symmetry properties*. In Section 1.1 we saw **Anonymity**, which expressed symmetry with respect

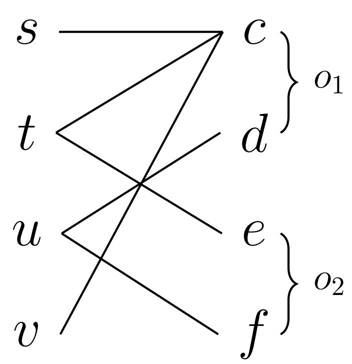


Figure 2.4: A network isomorphic to the one shown in Fig. 2.2.

to voters: if voters i and j swap their ballots, the election result remains the same. **Neutrality** expresses symmetry with respect to candidates: if we re-label candidates, the outcome remains the same up to re-labelling. In general, symmetry requires that the output of some process depends only on *structural* features of the input, not the specific "names" of the entities involved.

For truth discovery, we can consider symmetry with respect to sources, objects and claims. The central concept is an *isomorphism* between networks.

Definition 2.3.2. *An* isomorphism *between networks N and N' is mapping $F : S \cup O \cup C \to S' \cup O' \cup C'$ such that*

1. *$F|_S$, $F|_O$ and $F|_C$ are bijections $S \to S'$, $O \to O'$ and $C \to C'$, respectively.*
2. *For all $s \in S$ and $c \in C$: $(s, c) \in R$ iff $(F(s), F(c)) \in R'$.*
3. *For all $c \in C$, $\text{obj}(F(c)) = F(\text{obj}(c))$.*

That is, F is a one-to-one correspondence between the sources, objects and claims of N and their N' counterparts, which respects the structure of the network. One can easily check that we also have $F(\text{src}_N(c)) = \text{src}_{N'}(F(c))$ and $F(\text{cl}_N(s)) = \text{cl}_{N'}(F(s))$. The symmetry axiom says an operator should not distinguish isomorphic networks.

> **Symmetry**. If F is an isomorphism between N and N', then $s \sqsubseteq_N^T s'$ iff $F(s) \sqsubseteq_{N'}^T F(s')$ and $c \preceq_N^T c'$ iff $F(c) \preceq_{N'}^T F(c')$.

We illustrate **Symmetry** with an example.

Example 2.3.2. *Consider the network N from Fig. 2.2 and N' from Fig. 2.4, where we take the sources, objects and domains to be the same in both networks. Then N and N' are isomorphic via the mapping F expressed in cycle notation as $(suv)(cf)(de)(o_1o_2)$. For example, s plays the same role in N as u in N', c plays the same role in N as f in N', the role of objects o_1 and o_2 are swapped, etc.* ***Symmetry*** *requires that the source and claim rankings in N' are already* *determined by the rankings of* N*. For example, if the source ranking in* N *is* $s \sqsubset_N u \simeq_N v \sqsubset_N t$*, we must have* $u \sqsubset_{N'} v \simeq_{N'} s \sqsubset_{N'} t$*.*

An *automorphism* is an isomorphism F from a network N to itself. For example, F which swaps u and v in N from Fig. 2.2 is an automorphism, since u and v play exactly the same role in N. **Symmetry** implies that $u \simeq v$, and in fact this holds more generally.

Proposition 2.3.2. *If* F *is an automorphism on* N *and* T *satisfies* ***Symmetry****, then* $s \simeq_N^T F(s)$ *and* $c \approx_N^T F(c)$*, for all* $s \in S$ *and* $c \in C$*.*

Proof. We show $s \simeq_N^T F(s)$ for all sources s; the result for claims is similar. Take $s \in S$. Since S is finite and F restricts to a bijection $S \to S$, an argument identical to the one in the proof of Lemma 2.3.1 shows there is some $k \geq 1$ such that $s = F^k(s)$.

First suppose $s \sqsubseteq_N^T F(s)$. By **Symmetry** we may apply F to both sides; doing so repeatedly yields $F^n(s) \sqsubseteq_N^T F^{n+1}(s)$ for all $n \geq 1$. By transitivity of $\sqsubseteq_N^T$, we get $F(s) \sqsubseteq_N^T F^n(s)$. Taking $n = k$ gives $F(s) \sqsubseteq_N^T F^k(s) = s$, so $s \simeq_N^T F(s)$.

Now suppose $F(s) \sqsubseteq_N^T s$. By an identical argument, $F^n(s) \sqsubseteq_N^T F(s)$ for all $n \geq 1$; taking $n = k$ gives $s \sqsubseteq_N^T F(s)$, so $s \simeq_N^T F(s)$ again.

Since $\sqsubseteq_N^T$ is total these cases are exhaustive, and we are done. □

Proposition 2.3.2 is useful for showing certain sources and claims must rank equally. For example, take the network N from Fig. 2.3. Intuitively this network displays internal symmetry within the sources for each claim and between the claims themselves. Indeed, the functions $F = (st)(uv)$ and $G = (su)(tv)(cd)$ are automorphisms. By Proposition 2.3.2, any operator T satisfying **Symmetry** must output flat rankings $s \simeq t \simeq u \simeq v$ and $c \approx d$.

2.3.3 Monotonicity

Given that voting is not a viable truth discovery method, the believability of a claim c should not increase monotonically with $|\operatorname{src}_N(c)|$. Moreover, it should not increase with the *set* of sources $\operatorname{src}_N(c)$, ordered by set inclusion: $\operatorname{src}_N(c) \subseteq \operatorname{src}_N(d)$ should not in general imply $c \preceq d$. Indeed, consider an adversarial source t deliberately making false claims, and suppose $\operatorname{src}_N(c) = \{s\}$ and $\operatorname{src}_N(d) = \{s, t\}$. Then $\operatorname{src}_N(c) \subseteq \operatorname{src}_N(d)$, but the extra support from t should actually *decrease* the believability of d – since t only provides false claims – not increase it.

Nevertheless, there is a sense in which – all else being equal – a claim with more sources is more believable. The above examples show that some subtlety is needed in formulating this as a general principle, and that trust should be taken into account in doing so.

In this section we consider monotonicity properties of two kinds: monotonicity *within* a network, and monotonicity *between* networks as more reports are added. We start with the latter by adapting the idea of *positive responsiveness* from social choice theory.

Responsiveness. In the context of voting, positive responsiveness requires that a voter changing their ballot to prefer a winning candidate A does not cause A to fare worse in the collective ranking (Zwicker 2016) A naive version of positive responsiveness for truth discovery says that if we change a network N by adding a new report (s, c) – possibly removing reports from s conflicting with c – then c should move strictly up in the claim ranking. Clearly this neglects to consider the trustworthiness of s, and is thus an undesirable property (e.g. consider s adversarial as described above). Our first monotonicity axiom weakens this naive property by only considering "fresh" sources s not providing any reports in the original network N. Intuitively, we have no reason to believe such sources are untrustworthy, and they should therefore have a positive effect when making a claim. In what follows, when $\mathrm{cl}_N(s) = \emptyset$ we write $N + (s, c)$ for the network $(S, O, D, R \cup \{(s, c)\})$.

> **Fresh-pos-resp**. Suppose $\mathrm{cl}_N(s) = \emptyset$. Then for all $c \in C$ and $d \in C \setminus \{c\}$, $d \preceq^T_N c$ implies $d \prec^T_{N+(s,c)} c$.

That is, if c was already at least as believable as d, then a fresh report makes c *strictly* more believable in the new network.[5] What about the effects of a fresh report for c on source trustworthiness? According to the mutual dependence between the source and claim rankings – captured in a static network via the coherence properties – sources already claiming c should become more trusted, whereas those claiming a conflicting claim d should become less trusted.

> **Source-pos-resp**. Suppose $s \in \mathrm{antisrc}_N(c)$, $t \in \mathrm{src}_N(c)$, and $\mathrm{cl}_N(u) = \emptyset$. Then $s \sqsubseteq^T_N t$ implies $s \sqsubset^T_{N+(u,c)} t$.

Note that **Source-pos-resp** does not say anything about the ranking of the fresh source u. We consider another example.

Example 2.3.3. *Fig. 2.5 illustrates* ***Fresh-pos-resp*** *and* ***Source-pos-resp****. Let N_0 denote the network including only the solid edges, $N_1 = N_0 + (u, f)$, and $N_2 = N_1 + (v, f)$. Note that N_2 is our running example network from Fig. 2.2. Assuming* ***Symmetry****, everything is tied in N_0: we have $s \simeq_{N_0} t$ and $c \approx_{N_0} d \approx_{N_0} \approx_{N_0} e \approx_{N_0} f$. Since N_1 is the result of adding the report (u, f) and u makes no claims in N_0,* ***Fresh-pos-resp*** *gives $e \prec_{N_1} f$. Since $s \in \mathrm{src}_{N_0}(e) \subseteq \mathrm{antisrc}_{N_0}(f)$ and $t \in \mathrm{src}_{N_0}(f)$,* ***Source-pos-resp*** *gives $s \sqsubset_{N_1} t$. Going from N_1 to N_2 we can repeat exactly the same arguments to find $e \prec_{N_2} f$ and $s \sqsubset_{N_2} t$.*

[5] Note that N and $N + (s, c)$ share the same set of objects O and domains D, so the set of possible claims in both networks are the same. Consequently we are justified in treating c and d as claims in both networks.

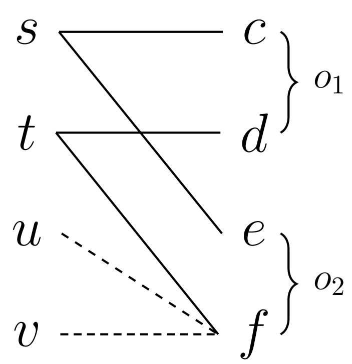


*Figure 2.5: Networks N_0 (solid edges only), $N_1 = N_0 + (u, f)$ and $N_2 = N_1 + (v, f)$ illustrating **Fresh-pos-resp** and **Source-pos-resp**.*

*Bringing **Claim-coherence** in too, $s \sqsubset_{N_2} t$ gives $c \prec_{N_2} d$. Thus, **Claim-coherence**, **Symmetry**, **Fresh-pos-resp** and **Source-pos-resp** are enough to capture our intuitions about this network as described in the introduction.*

In the special case where a network contains reports only for a single object, the responsiveness properties and **Symmetry** actually force an operator to rank claims by voting, and to rank sources by the vote count of their claims. Note that each source provides at most one report in this case, by condition (2) in the definition of a network. Consequently there is little structure in such networks, as we cannot look at how sources interact over multiple objects to determine trustworthiness. We therefore argue that voting is reasonable behaviour in this special case.

Proposition 2.3.3. *Suppose there is $o \in O$ such that $\mathrm{src}_N(o') = \emptyset$ for all $o \neq o'$. Then*

1. *If T satisfies **Symmetry** and **Fresh-pos-resp**, then for all $c, d \in \mathrm{cl}_N(o)$:*

$$c \preceq_N^T d \iff |\,\mathrm{src}_N(c)| \leq |\,\mathrm{src}_N(d)|.$$

2. *If T satisfies **Symmetry** and **Source-pos-resp**, then for all $s, t \in S$ with $\mathrm{cl}_N(s), \mathrm{cl}_N(t) \neq \emptyset$,*

$$s \sqsubseteq_N^T t \iff |\,\mathrm{src}_N(c_s)| \leq |\,\mathrm{src}_N(c_t)|,$$

where c_s and c_t are the unique claims reported by s and t respectively.

While Proposition 2.3.3 only addresses a somewhat trivial case, it will turn out to be useful in characterising voting behaviour more generally in Sections 2.3.4 and 2.3.6. It can be seen as one of the many generalisations of *May's Theorem* (May 1952), which characterises the majority voting rule in two-candidate elections. To prove it, we need a preliminary result.

Lemma 2.3.2. *Suppose* $|\operatorname{src}_N(c)| = |\operatorname{src}_N(d)|$, $\operatorname{obj}(c) = \operatorname{obj}(d)$, *and for all* $s \in \operatorname{src}_N(c) \cup \operatorname{src}_N(d)$, $|\operatorname{cl}_N(s)| = 1$. *Then for any operator* T *satisfying* ***Symmetry***, $c \approx_N^T d$.

Proof. Without loss of generality, assume $c \neq d$. Since $\operatorname{obj}(c) = \operatorname{obj}(d)$, we have $c \in \operatorname{conflict}_N(d)$ and thus $\operatorname{src}_N(c) \cap \operatorname{src}_N(d) = \emptyset$. Since $|\operatorname{src}_N(c)| = |\operatorname{src}_N(d)|$ there exists a bijection $\hat{\varphi} : \operatorname{src}_N(c) \to \operatorname{src}_N(d)$. We extend this to a bijection $\varphi : S \to S$ by

$$\varphi(s) = \begin{cases} \hat{\varphi}(s), & s \in \operatorname{src}_N(c) \\ \hat{\varphi}^{-1}(s), & s \in \operatorname{src}_N(d) \\ s, & \text{otherwise.} \end{cases}$$

Now let $F : S \cup C \cup O \to S \cup C \cup O$ be defined by $F|_S = \varphi$, $F|_C = (cd)$ and $f|_O = \mathrm{id}$. That is, F permutes sources according to φ, swaps claims c and d, and leaves objects as they are. Since $F(c) = d$, to show $c \approx_N^T d$ it is sufficient by Proposition 2.3.2 to show that F is an automorphism on N.

It is easily seen that the restrictions of F to S, C and O respectively, are bijective. Moreover, we have $\operatorname{obj}(F(e)) = F(\operatorname{obj}(e))$ for all claims e since $F(o) = o$ and $\operatorname{obj}(c) = \operatorname{obj}(d)$. It remains to show that $(s, e) \in R$ iff $(F(s), F(e)) \in R$.

For the left-to-right direction, suppose $(s, e) \in R$. First suppose $s \in \operatorname{src}_N(c)$. Then $F(s) = \hat{\varphi}(s) \in \operatorname{src}_N(d)$, so $(F(s), d) \in R$. By assumption we have $|\operatorname{cl}_N(s)| = 1$, so in fact c is the unique claim reported by s. Thus $e = c$. Consequently

$$(F(s), F(e)) = (F(s), d) \in R$$

as required. The case for $s \in \operatorname{src}_N(d)$ follows by a near-identical argument. Finally, if $s \notin \operatorname{src}_N(c) \cup \operatorname{src}_N(d)$ then $F(s) = s$ and $e \notin \{c, d\}$, so $F(e) = e$. Thus $(F(s), F(e)) = (s, e) \in R$.

For the right-to-left direction, suppose $(F(s), F(e)) \in R$. Applying the argument above we have $(F^2(s), F^2(e)) \in R$ also. But note that $F = F^{-1}$, so $F^2 = \mathrm{id}$. Hence $(s, e) \in R$, as required. This completes the proof. ☐

Proof of Proposition 2.3.3. We prove (1) only, since (2) can be shown using essentially the same argument with **Source-pos-resp** taking the place of **Fresh-pos-resp**.

Suppose T satisfies **Symmetry** and **Fresh-pos-resp**, and take N as stated in Proposition 2.3.3. It is sufficient to show that, for all $c, d \in \operatorname{cl}_N(o)$,

$$|\operatorname{src}_N(c)| \leq |\operatorname{src}_N(d)| \implies c \preceq_N^T d \tag{2.4}$$

$$|\operatorname{src}_N(c)| < |\operatorname{src}_N(d)| \implies c \prec_N^T d. \tag{2.5}$$

First we show (2.4). Suppose $|\operatorname{src}_N(c)| \leq |\operatorname{src}_N(d)|$. Assume without loss of generality that $c \neq d$. Write $k = |\operatorname{src}_N(d)| - |\operatorname{src}_N(c)| \geq 0$. Let $X =$

$\{s_1, \ldots, s_k\}$ be an arbitrary subset of $\mathrm{src}_N(d)$ of size k. Let N_0 denote the network in which all claims from sources in X are removed. Note that since N does not contain reports for objects other than o, by the consistency property (2) in Definition 2.1.1 we have that sources in X *only* report d. We construct networks $N_1, \ldots, N_k$ in which these claims are added back in: for $0 \le i \le k-1$, set

$$N_{i+1} = N_i + (s_{i+1}, d).$$

Then N_k is just the original network N. Note that $\mathrm{cl}_{N_i}(s_j) = \emptyset$ for $j > i$. Next we show by induction that for all $0 \le i \le k$,

$$c \preceq^T_{N_i} d, \text{ and if } i > 0 \text{ then } c \prec^T_{N_i} d. \tag{2.6}$$

For the base case $i = 0$, note that since only reports for d were removed in constructing N_0, we have $\mathrm{src}_{N_0}(c) = \mathrm{src}_N(c)$. Consequently,

$$|\,\mathrm{src}_{N_0}(d)| = |\,\mathrm{src}_N(d) \setminus X| = |\,\mathrm{src}_N(d)| - k = |\,\mathrm{src}_N(c)| = |\,\mathrm{src}_{N_0}(c)|.$$

Note also that $\mathrm{obj}(c) = \mathrm{obj}(d)$ – since by assumption $c, d \in \mathrm{cl}_N(o)$ – and for $s \in \mathrm{src}_{N_0}(c) \cup \mathrm{src}_{N_0}(d)$ we have $|\,\mathrm{cl}_{N_0}(s)| = 1$ since N_0 also only contains reports for o. The hypothesis of Lemma 2.3.2 are satisfied, so we have $c \approx^T_{N_0} d$. In particular, $c \preceq^T_{N_0} d$ as required.

Now for the inductive step, suppose (2.6) holds for i. Since $\mathrm{cl}_{N_i}(s_{i+1}) = \emptyset$, **Fresh-pos-resp** and the inductive hypothesis give $c \prec^T_{N_{i+1}} d$, as required.

Finally, (2.4) follows by taking $i = k$ in (2.6), recalling that $N = N_k$. Moreover, (2.5) follows by exactly the same argument, noting that when $|\,\mathrm{src}_N(c)| < |\,\mathrm{src}_N(d)|$ we have $k > 0$, so $c \prec^T_{N_k} d$ by (2.6) again. □

Trust-based monotonicity. Suppose $\mathrm{src}_N(d) = \mathrm{src}_N(c) \cup \{s\}$. The relative ranking of c and d depends on the marginal effect of s: if s is "trustworthy" then d gains credibility from the extra support of s, whereas if s is "untrustworthy" this extra support has the opposite effect. Our next axiom requires that such marginal effects are compatible with the source trustworthiness ranking. First, some terminology is required.

Definition 2.3.3. *Given a network N, a source $s \in S$ is* marginally trustworthy *with respect to an operator T if there exist claims $c, d \in C$ such that $s \notin \mathrm{src}_N(c)$, $\mathrm{src}_N(d) = \mathrm{src}_N(c) \cup \{s\}$ and $c \preceq^T_N d$. Similarly, s is* marginally untrustworthy *if there are $c, d \in C$ such that $s \notin \mathrm{src}_N(c)$, $\mathrm{src}_N(d) = \mathrm{src}_N(c) \cup \{s\}$ and $d \preceq^T_N c$.*

These properties express something about the trustworthiness of sources via the *claim* ranking $\preceq^T_N$, akin to how **Source-coherence** looks at trustworthiness via the claims reported by a source. Note that it is possible for a source to be both marginally trustworthy and untrustworthy. Naturally, marginally untrustworthy sources should rank lower than marginally trustworthy ones.

Marginal-trustworthiness. If s is marginally untrustworthy and t is marginally trustworthy, then $s \sqsubseteq_N^T t$.

Equipped with a notion of marginal trustworthiness, we can also state a trust-aware monotonicity axiom for claims.

Trust-based-monotonicity. Suppose $\mathrm{src}_N(d) = \mathrm{src}_N(c) \cup Z$, where $\mathrm{src}_N(c) \cap Z = \emptyset$. Then

1. If each $s \in Z$ is marginally trustworthy, $c \preceq_N^T d$.
2. If each $s \in Z$ is marginally untrustworthy, $d \preceq_N^T c$.

Informally, **Trust-based-monotonicity** says that if each $s \in Z$ has a positive (or at least, not negative) impact on some claim in N, as measured by $\preceq_N^T$, then the sources in Z acting collectively should also have a positive impact. Also note that in the case $Z = \{s\}$, **Trust-based-monotonicity** implies that the marginal impact of s is consistent across the network.

2.3.4 Independence

Another common class of axioms in social choice theory are *independence* axioms, which require that some aspect of the output is independent of "irrelevant" parts of the input. The original example is Arrow's *Independence of Irrelevant Alternatives* (IIA) in voting theory (Arrow 1952), which says, roughly speaking, that the ranking of candidates A and B should depend only on the individual rankings of A and B, not on any "irrelevant" alternative C. It has been adapted to several settings in which the axiomatic method has been applied. Perhaps closest to our setting is judgement aggregation, where independence requires the collective acceptance of a report φ does not depend on how the individuals accept or reject some other report ψ (Endriss 2016).

A version of IIA can be easily stated in our framework: the ranking of claims c and d should depend only on the sources reporting c and d, not on the sources for other claims. However, this axiom is clearly *undesirable* for truth discovery. Indeed, consider again the network N from Fig. 2.2. As we have argued informally, claim c is intuitively weaker than d because of the way in which their respective sources interact with other claims in the network. Nevertheless, we state this axiom as a point of comparison with classical social choice problems such as voting.

Classical-independence. Suppose $C_N = C_{N'}$. Then $\mathrm{src}_N(c) = \mathrm{src}_{N'}(c)$ and $\mathrm{src}_N(d) = \mathrm{src}_{N'}(d)$ implies $c \preceq_N^T d$ iff $c \preceq_{N'}^T d$.

That is, if c and d have the same sources in N and N', they have the same relative ranking in both networks. The undesirability of **Classical-independence** can be formalised axiomatically: together with our earlier axioms, it implies voting-like behaviour within the claims for each object.[6]

Note that for the special case of binary networks, similar results have been shown in the literature on binary aggregation with abstentions (Christoff and Grossi 2017).

Proposition 2.3.4. *Suppose T satisfies **Symmetry**, **Fresh-pos-resp** and **Classical-independence**. Then for all $o \in O$ and $c, d \in \mathrm{cl}_N(o)$,*

$$c \preceq_N^T d \iff |\mathrm{src}_N(c)| \leq |\mathrm{src}_N(d)|.$$

Proof. Take $c, d \in \mathrm{cl}_N(o)$. Let the network N' have the same sources, objects and domains as N, but with reports $R' = R \cap (S \times \{c, d\})$. That is, N' discards all reports for claims other than c and d. Then we have $\mathrm{src}_{N'}(c) = \mathrm{src}_N(c)$, $\mathrm{src}_{N'}(d) = \mathrm{src}_N(d)$, and $\mathrm{src}_{N'}(e) = \emptyset$ for all $e \notin \{c, d\}$. By **Classical-independence**, $c \preceq_N^T d$ iff $c \preceq_{N'}^T d$.

Now, note that since $c, d \in \mathrm{cl}_N(o)$, for $o' \neq o$ and $e \in \mathrm{cl}_N(o')$ we have $e \notin \{c, d\}$, so $\mathrm{src}_N(e) = \emptyset$. Hence $\mathrm{src}_N(o') = \emptyset$ for such o'. Since T satisfies **Symmetry** and **Fresh-pos-resp**, we may apply Proposition 2.3.3 (1) to find $c \preceq_{N'}^T d$ iff $|\mathrm{src}_{N'}(c)| \leq |\mathrm{src}_{N'}(d)|$. But $|\mathrm{src}_{N'}(c)| = |\mathrm{src}_N(c)|$, and likewise for d. Consequently

$$c \preceq_N^T d \iff c \preceq_{N'}^T d \iff |\mathrm{src}_{N'}(c)| \leq |\mathrm{src}_{N'}(d)| \iff |\mathrm{src}_N(c)| \leq |\mathrm{src}_N(d)|$$

as desired. □

While this result appears similar to Proposition 2.3.3, the crucial difference is that we no longer restrict to the case sources only report on a single object, where voting is justified. This is the (overly strong) role **Classical-independence** plays: it allows the complexity of a multi-object network to be reduced to a single-object network, where the ranking trivialises.

Recalling from Example 2.3.3 that **Claim-coherence**, **Symmetry**, **Fresh-pos-resp** and **Source-pos-resp** are enough to ensure $c \prec d$ in our running example network from Fig. 2.2 (whereas per-object voting gives $c \approx d$), we obtain an impossibility result with **Classical-independence**. In fact we obtain *two* impossibility results, since **Source-pos-resp** can also be replaced with **Source-coherence**.

Theorem 2.3.1. *Suppose an operator satisfies **Symmetry**, **Claim-coherence** and **Fresh-pos-resp**. Then the following axioms cannot hold simultaneously.*

1. ***Source-pos-resp*** *and* ***Classical-independence***.
2. ***Source-coherence*** *and* ***Classical-independence***.

Proof.

[6] We give a further axiom which implies voting behaviour for claims of *different* objects – and leads to a complete characterisation of voting – in Section 2.3.6.

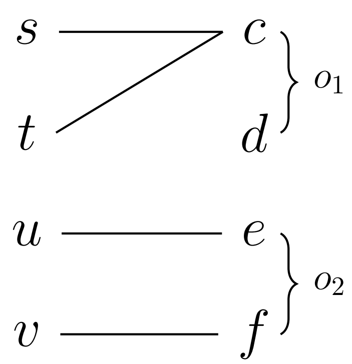


Figure 2.6: A network illustrating ***Disjoint-independence****.*

1. The impossibility of these axioms holding together follows from Example 2.3.3 and Proposition 2.3.4, as described above.

2. Let N be as shown in Fig. 2.2. Suppose some operator T satisfies the stated axioms. From Proposition 2.3.4 we get $c \approx^T_N d$ and $e \prec^T_N f$. Considering sources s and t, **Source-coherence** gives $s \sqsubset^T_N t$. But now **Claim-coherence** gives $c \prec^T_N d$: contradiction.

$\square$

By only looking at a claim's sources, **Classical-independence** ignores the indirect interaction with other sources and claims in the network. Our next axiom accounts for such interactions by considering networks with *disjoint sub-networks*, such as the one shown in Fig. 2.6. Intuitively, while the sources and claims within a sub-network may interact in complex ways, the fact that the sub-networks have no sources or objects in common means there is no interaction *between* them. Accordingly, the ranking for one should not depend on the other. We formalise this by considering unions of *disjoint networks*.[7]

Definition 2.3.4. *Networks N and N' are* disjoint *if $S \cap S' = \emptyset$ and $O \cap O' = \emptyset$. For N, N' disjoint, their* union *is the network $N \sqcup N' = (S \cup S', O \cup O', \hat{D}, R \cup R')$, where $\hat{D}_o = D_o$ for $o \in O$, and $\hat{D}_o = D'_o$ for $o \in O'$.*

Note that if N and N' are disjoint, it follows that $C \cap C' = \emptyset$ also. The following axiom says that the ranking of sources and claims is unaffected by the addition of a disjoint network.

> **Disjoint-independence**. If N and N' are disjoint, $s, t \in S$, and $c, d \in C$, then $s \sqsubseteq^T_N t$ iff $s \sqsubseteq^T_{N \sqcup N'} t$ and $c \preceq^T_N d$ iff $c \preceq^T_{N \sqcup N'} d$.

[7] Note that it is possible to define the disjoint union of an arbitrary collection of (not necessarily disjoint) networks in a manner similar to the disjoint union of a collection of sets $\bigsqcup_{i \in I} X_i$, but we do not need this generality here.

Disjoint-independence also has a natural graphical interpretation. For any network we may consider the tripartite graph with nodes $S \cup C \cup O$, and with edges from sources to their claims and from claims to their respective objects. Any network N can then be expressed as the disjoint union of the networks corresponding to the *connected components* of this graph. For instance, the network shown in Fig. 2.6 – viewed as a graph – has two connected components. In graphical terms, **Disjoint-independence** says that the ranking of sources and claims in one component does not depend on the presence of the other.

2.3.5 Conflicting claims

Our axioms so far have not made use of the conflict relation between claims. Intuitively, distinct claims c, c' for the same object o cannot both be true, so belief in c should come at the expense of belief in c'. Similarly, if the antisources of c – that is, the sources who report claims conflicting with c – are seen as less trustworthy than the antisources of c', then the attack on c is less damaging than that of c', so c should be more believable than c'. Note that these are again coherence principles, which constrain how the claim ranking $\preceq$ coheres with both the source ranking $\sqsubseteq$ and the conflict relation. We formulate them as axioms.

> **Conflict-coherence**. If $\mathrm{conflict}_N(c)$ strictly precedes $\mathrm{conflict}_N(c')$ pairwise with respect to $\preceq_N^T$, then $c' \prec_N^T c$.
>
> **Anti-coherence**. If $\mathrm{antisrc}_N(c)$ strictly precedes $\mathrm{antisrc}_N(c')$ pairwise with respect to $\sqsubseteq_N^T$, then $c' \prec_N^T c$.

While both **Conflict-coherence** and **Anti-coherence** appear reasonable in isolation, there is an inherent tension between them and our earlier coherence axioms. Together with symmetry and responsiveness axioms, we have an impossibility result.

Theorem 2.3.2. *Suppose an operator satisfies* ***Symmetry*** *and* ***Claim-coherence****. Then the following axioms cannot hold simultaneously.*

1. ***Fresh-pos-resp****,* ***Source-coherence*** *and* ***Conflict-coherence****,*
2. ***Source-pos-resp*** *and* ***Conflict-coherence****.*
3. ***Source-pos-resp*** *and* ***Anti-coherence****.*

Proof. Suppose T satisfies **Symmetry** and **Claim-coherence**. Throughout the proof, let N_0 denote the network shown in Fig. 2.7 excluding the dashed edge, and let $N_1 = N + (u, f)$ denote the network including the dashed edge. We first note some consequences of the axioms in both networks. In N_0, the mapping $(s\ s')(t\ t')(c\ c')(d\ d')(o\ o')(e\ f)$ is an automorphism, so we have

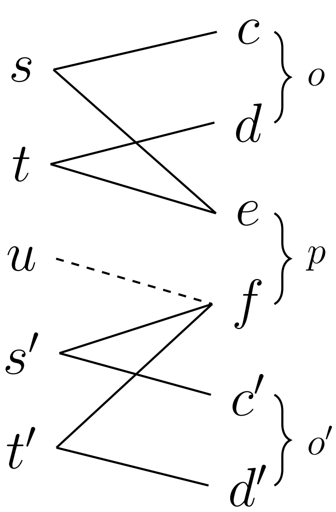


Figure 2.7: Network illustrating the impossibility results of Theorem 2.3.2.

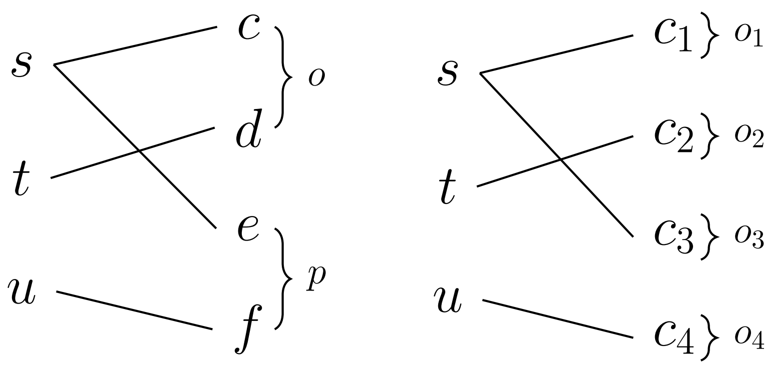
$s \simeq^T_{N_0} s'$ and $e \approx^T_{N_0} f$. Note that $\mathrm{src}_{N_0}(u) = \emptyset$, $s \in \mathrm{antisrc}_{N_0}(f)$ and $s' \in \mathrm{src}_{N_0}(f)$. If T additionally satisfies **Fresh-pos-resp**, we get $e \prec^T_{N_1} f$. If T instead satisfies **Source-pos-resp**, we get $s \sqsubset^T_{N_1} s'$. Considering N_1 alone, the mapping $(s\ t)(s'\ t')(c\ d)(c'\ d')$ is an automorphism, so **Symmetry** gives $c \approx^T_{N_1} d$ and $c' \approx^T_{N_1} d'$.

1. Suppose T also satisfies **Fresh-pos-resp**, **Source-coherence** and **Conflict-coherence**. First we claim $c \approx^T_{N_1} c'$. Indeed, suppose not. If $c' \prec^T_{N_1} c$, we may note that $\mathrm{conflict}_{N_1}(d) = \{c\}$ and $\mathrm{conflict}_{N_1}(d') = \{c'\}$, and apply **Conflict-coherence** to get $d \prec^T_{N_1} d'$. But by **Symmetry** as above, we have $c \approx^T_{N_1} d$ and $c' \approx^T_{N_1} d'$. Consequently $c \approx^T_{N_1} d \prec^T_{N_1} d' \approx^T_{N_1} c'$, i.e. $c \prec^T_{N_1} c'$. Clearly this contradicts $c' \prec^T_{N_1} c$. If $c \prec^T_{N_1} c'$ we obtain a contradiction by an identical argument. Hence $c \approx^T_{N_1} c'$.

   Now, by **Fresh-pos-resp** and **Symmetry** as noted above, we have $e \prec^T_{N_1} f$. **Source-coherence** for s and s' therefore gives $s \sqsubset^T_{N_1} s'$. But considering c and c', **Claim-coherence** gives $c \prec^T_{N_1} c'$. This contradicts $c \approx^T_{N_1} c'$, and we are done.

2. Suppose T additionally satisfies **Source-pos-resp** and **Conflict-coherence**. By the same argument as above, **Conflict-coherence** and **Symmetry** together dictate that $c \approx^T_{N_1} c'$. But by **Symmetry** and **Source-pos-resp**, we have $s \sqsubset^T_{N_1} s'$. **Claim-coherence** then implies $c \prec^T_{N_1} c'$: contradiction.

3. Suppose T additionally satisfies **Source-pos-resp** and **Anti-coherence**. Again, $s \sqsubset^T_{N_1} s'$. **Claim-coherence** implies $c \prec^T_{N_1} c'$. Since $\mathrm{antisrc}_{N_1}(d) = \{s\}$ and $\mathrm{antisrc}_{N_1}(d') = \{s'\}$, **Anti-coherence** gives $d' \prec^T_{N_1} d$. But recall that, by **Symmetry**, $c \approx^T_{N_1} d$ and $c' \approx^T_{N_1} d'$. Hence $c \prec^T_{N_1} c' \approx^T_{N_1} d' \prec^T_{N_1} d \approx^T_{N_1} c$, i.e. $c \prec^T_{N_1} c$: contradiction.



(a) N (b) N'

Figure 2.8: Illustration of an object reduction of a network.

□

Note that all four coherence axioms can be satisfied at the same time, e.g. by the trivial operator which outputs constant scores $T_N(s) = T_N(c) = 0$. Of course, this operator violates both **Fresh-pos-resp** and **Source-pos-resp**.

2.3.6 Axiomatic Characterisation of Voting

Recall from Proposition 2.3.4 that **Symmetry**, **Fresh-pos-resp** and **Classical-independence** force an operator to rank claims for the object simply by their number of sources, as in voting from Section 2.2.1. In this section we give two further axioms which force this ranking even for claims across different objects, and thus characterise T^{vote} completely. Like **Classical-independence**, these axioms are *not* desirable properties, and are introduced only to capture the behaviour of voting. The first axiom simply says that the source ranking is flat.

Flat-sources. For all $s, s' \in S$, $s \simeq_N^T s'$.

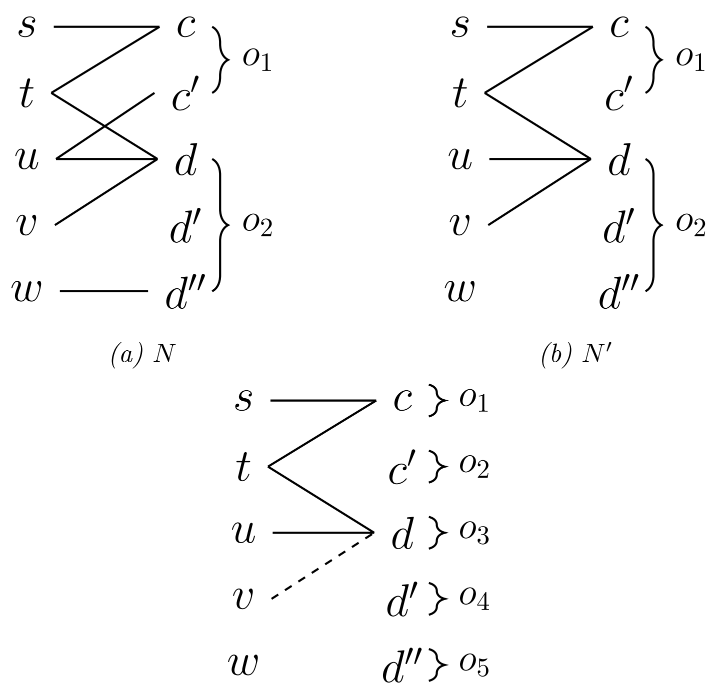
The second axiom says that objects play no role: it is only the relation between sources and claims which affects the rankings. That is, we can ignore the conflict relation between claims. To define the axiom we introduce a notion of "reducing" the objects of a network.

Definition 2.3.5. *A network N' is an* object reduction *of N via $f : C_N \to C_{N'}$ if*

1. *$S' = S$.*

2. *f is a bijection $C_N \to C_{N'}$ such that $(s, c) \in R$ iff $(s, f(c)) \in R'$.*

3. *For all $o \in O'$, $|D'_o| = 1$.*



(a) N (b) N'

(c) N'' (all edges) and N_0 (excluding dashed edge)

*Figure 2.9: Illustration of the proof of Theorem 2.3.3. In N', reports for claims other than c and d are removed. N'' is an object reduction of N'. The dashed edge shows the reports added when **Fresh-pos-resp** is applied.*

Note that every network N has an object reduction since the set of possible objects $\mathbb{O}$ is infinite; we may take O' to be any subset of $\mathbb{O}$ of size $|C_N|$, take $D'_o = \{v\}$ for some fixed $v \in \mathbb{V}$, and set R' accordingly. Fig. 2.8 shows an example of an object reduction. Note that the network N' has only a single claim for each object, and the structure of the reports – i.e. the edges shown in Fig. 2.8 – is the same in N and N'. Going from N to N' loses information about which claims conflict with one another, and our axioms in Section 2.3.5 explicitly require that this information *does* affects the rankings. Voting does not use this information, however, which leads to the following axiom.

> **Object-irrelevance**. If N' is an object reduction of N via f, then $c \preceq^T_N d$ iff $f(c) \preceq^T_N f(d)$.

Note that **Object-irrelevance** is similar in form to **Symmetry**, but rather than requiring rankings are invariant under isomorphisms – which preserve the relevant structure of a network – it requires rankings are invariant under object reductions.

We can now characterise voting, up to ranking equivalence.

Theorem 2.3.3. *An operator T satisfies **Symmetry**, **Fresh-pos-resp**, **Classical-independence**, **Flat-sources** and **Object-irrelevance** if and only if $T \sim T^{\text{vote}}$.*

Proof (sketch). The "if" direction is straightforward. For the "only if" direction, take an operator T with the stated axioms. **Flat-sources** immediately implies $\sqsubseteq_N^T = \sqsubseteq_N^{T^{\text{vote}}}$ for all networks N. For the claim rankings, we take a similar approach to the proof of Proposition 2.3.4 and only sketch the argument here. An illustration of the proof is shown in Fig. 2.9.

Take any network N and claims c, d. We first remove all reports for other claims to produce N'; this preserves rankings by **Classical-independence**. Taking N'' to be any object reduction of N', we ensure c and d are the only claims for their respective objects,[8] and rankings are again preserved by **Object-irrelevance**. As before, it suffices to show that $|\operatorname{src}_N(c)| \leq |\operatorname{src}_N(d)|$ implies $c \preceq_N^T d$ and $|\operatorname{src}_N(c)| < |\operatorname{src}_N(d)|$ implies $c \prec_N^T d$, since c and d are arbitrary.

Write $k = |\operatorname{src}_N(d)| - |\operatorname{src}_N(c)| \geq 0$. Choosing k sources from $\operatorname{src}_N(d) \setminus \operatorname{src}_N(c)$, let N_0 be the network obtained from N'' in which reports for d from these sources are removed. Note that such sources *only* report d, since reports for other claims were removed in the construction of N'. Then $|\operatorname{src}_{N_0}(c)| = |\operatorname{src}_{N_0}(d)|$. The fact that $|D''_{\operatorname{obj}(c)}| = |D''_{\operatorname{obj}(d)}| = 1$ ensures we are able to choose an automorphism on N_0 which swaps c and d (and swaps $\operatorname{src}_{N_0}(c) \setminus \operatorname{src}_{N_0}(d)$ with $\operatorname{src}_{N_0}(d) \setminus \operatorname{src}_{N_0}(c)$). By **Symmetry**, $c \approx_{N_0}^T d$.

If $k = 0$ then $N_0 = N''$, and we are done. Otherwise, by repeated applications of **Fresh-pos-resp** we may add the removed reports back in to N_0 to get $c \prec_{N''}^T d$. Since claim rankings are the same in N'' as in N, this completes the proof. □

2.4 Satisfaction of the Axioms

In the previous section we introduced several axioms for truth discovery. We now turn back to some of the example operators from Section 2.2, to assess which axioms hold for each operator. For simplicity we skip TruthFinder and CRH, which due to their somewhat complicated form are not straightforward to analyse. The results are summarised in Table 2.2.

Weighted Voting. First we consider weighted voting. The following axioms hold for *any* choice of weighting w.

Lemma 2.4.1. *Let w be a weighting. Then T^w satisfies **Claim-coherence**, **Marginal-trustworthiness** and **Trust-based-monotonicity**.*

Proof. **Claim-coherence** follows easily using the definition of weighted voting and Proposition 2.3.1.

[8] Strictly speaking, we should define an object reduction f between N' and N'', and refer to $f(c)$ and $f(d)$ in N'' instead of c and d. For simplicity we identify c with $f(c)$ and d with $f(d)$ in this proof sketch.

Table 2.2: Axiom satisfaction for the example operators.

	Voting	WeightedAgg	Sums	USums
Claim-coherence	✓	✓	✓	✓
Source-coherence	X	X	✓	✓
Symmetry	✓	✓	✓	✓
Fresh-pos-resp	✓	✓	X	?
Source-pos-resp	X	✓	X	?
Marginal-trustworthiness	✓	✓	✓	✓
Trust-based-monotonicity	✓	✓	✓	✓
Classical-independence	✓	X	X	X
Disjoint-independence	✓	✓	X	X
Conflict-coherence	X	X	X	X
Anti-coherence	X	X	X	X

One can easily show that if s is marginally trustworthy with respect to T^w then $w_N(s) \geq 0$, and if s is marginally untrustworthy with respect to T^w then $w_N(s) \geq 0$, and **Marginal-trustworthiness** follows.

Finally, for **Trust-based-monotonicity** suppose $\mathrm{src}_N(d) = \mathrm{src}_N(c) \cup Z$, where $\mathrm{src}_N(c) \cap Z = \emptyset$. Then $T^w_N(d) = T^w_N(c) + \sum_{s \in Z} w_N(s)$. If each $s \in Z$ is marginally trustworthy then each $w_N(s)$ is non-negative, and so too is the sum. Hence $T^w_N(d) \geq T^w(c)$, so $c \preceq^{T^w}_N d$. If each $s \in Z$ is marginally untrustworthy then each $w_N(s)$ is non-positive, and similarly we get $d \preceq^{T^w}_N c$ as required. □

Corollary 2.4.1. *Any weightable operator satisfies **Claim-coherence**, **Marginal-trustworthiness** and **Trust-based-monotonicity**.*

Proof. This follows directly from Lemma 2.4.1 since each axiom only refers to ordinal properties of operators. □

Voting arises via the uniform weighting $w_N \equiv 1$. We have the following.

Theorem 2.4.1. *Voting satisfies **Claim-coherence**, **Symmetry**, **Fresh-pos-resp**, **Marginal-trustworthiness**, **Trust-based-monotonicity**, **Classical-independence** and **Disjoint-independence**. It does not satisfy **Source-coherence**, **Source-pos-resp**, **Conflict-coherence** or **Anti-coherence**.*

The proof is for the most part straightforward, and is omitted for brevity. For the particular choice of w for Weighted Agreement from Example 2.2.1, we have the following.

Theorem 2.4.2. *Weighted Agreement satisfies **Claim-coherence**, **Symmetry**, **Fresh-pos-resp**, **Source-pos-resp**, **Marginal-trustworthiness**, **Trust-based-monotonicity** and **Disjoint-independence**. It does not satisfy **Source-coherence**, **Classical-independence**, **Conflict-coherence** or **Anti-coherence**.*

Proof. For brevity, let w denote w^{agg} and T denote $T^{w^{\mathrm{agg}}}$. **Claim-coherence**, **Marginal-trustworthiness** and **Trust-based-monotonicity** follow from Lemma 2.4.1.

For **Symmetry**, suppose F is an isomorphism between networks N and N'. From the definition of an isomorphism we have $(s, c) \in R$ iff $(F(s), F(c)) \in R'$. Consequently $\mathrm{src}_N(c) = \{F^{-1}(s') \mid s' \in \mathrm{src}_{N'}(F(c))\}$ and $\mathrm{cl}_N(s) = \{F^{-1}(c') \mid c' \in \mathrm{cl}_{N'}(F(s))\}$. From this one can show $w_N(s) = w_{N'}(F(s))$, which then implies $T_N(s) = T_{N'}(F(s))$ and $T_N(c) = T_{N'}(F(c))$. **Symmetry** now follows.

For **Fresh-pos-resp** and **Source-pos-resp**, we use the following auxiliary result.

Claim 2.4.1. *Suppose* $\mathrm{cl}_N(u) = \emptyset$ *and let* c *be a claim. Then for all* $s \neq u$ *with* $\mathrm{cl}_N(s) \neq \emptyset$,

$$w_{N+(u,c)}(s) = w_N(s) + \frac{\mathbb{1}[c \in \mathrm{cl}_N(s)]}{|\,\mathrm{cl}_N(s)|}.$$

Proof of claim. First, note that for any claim d,

$$|\,\mathrm{src}_{N+(u,c)}(d)| = |\,\mathrm{src}_N(d)| + \mathbb{1}[c = d],$$

and since $s \neq u$ we have $\mathrm{cl}_{N+(u,c)}(s) = \mathrm{cl}_N(s)$. Consequently

$$\begin{aligned}
w_{N+(u,c)}(s) &= \sum_{d \in \mathrm{cl}_{N+(u,c)}(s)} \frac{|\,\mathrm{src}_{N+(u,c)}(d)|}{|\,\mathrm{cl}_{N+(u,c)}(s)|} \\
&= \sum_{d \in \mathrm{cl}_N(s)} \frac{|\,\mathrm{src}_N(d)| + \mathbb{1}[c = d]}{|\,\mathrm{cl}_N(s)|} \\
&= \underbrace{\sum_{d \in \mathrm{cl}_N(s)} \frac{|\,\mathrm{src}_N(d)|}{|\,\mathrm{cl}_N(s)|}}_{= w_N(s)} + \sum_{d \in \mathrm{cl}_N(s)} \underbrace{\frac{\mathbb{1}[c = d]}{|\,\mathrm{cl}_N(s)|}}_{=0 \text{ unless } c = d} \\
&= w_N(s) + \frac{\mathbb{1}[c \in \mathrm{cl}_N(s)]}{|\,\mathrm{cl}_N(s)|}.
\end{aligned}$$

$\diamondsuit$

Now, for **Fresh-pos-resp**, suppose $\mathrm{cl}_N(u) = \emptyset$, $c \neq d$ and $d \preceq_N^T c$. We need to show $d \prec^T_{N+(u,c)} c$. Indeed, using Claim 2.4.1 we have

$$
\begin{aligned}
&T_{N+(u,c)}(c) - T_{N+(u,c)}(d) = w_{N+(u,c)}(u) + \sum_{s \in \mathrm{src}_N(c)} w_{N+(u,c)}(s) - \sum_{s \in \mathrm{src}_N(d)} w_{N+(u,c)}(s) \\
&= |\,\mathrm{src}_N(c)| + 1 + \sum_{s \in \mathrm{src}_N(c)} \left(w_N(s) + \frac{1}{|\,\mathrm{cl}_N(s)|} \right) - \sum_{s \in \mathrm{src}_N(d)} \left(w_N(s) + \frac{\mathbb{1}[c \in \mathrm{cl}_N(s)]}{|\,\mathrm{cl}_N(s)|} \right) \\
&= |\,\mathrm{src}_N(c)| + 1 + T_N(c) + \sum_{s \in \mathrm{src}_N(c)} \frac{1}{|\,\mathrm{cl}_N(s)|} - T_N(d) - \sum_{s \in \mathrm{src}_N(c) \cap \mathrm{src}_N(d)} \frac{1}{|\,\mathrm{cl}_N(s)|} \\
&= |\,\mathrm{src}_N(c)| + 1 + \underbrace{T_N(c) - T_N(d)}_{\geq 0} + \sum_{s \in \mathrm{src}_N(c) \setminus \mathrm{src}_N(d)} \frac{1}{|\,\mathrm{cl}_N(s)|} \\
&\geq 1 > 0.
\end{aligned}
$$

This shows $T_{N+(u,c)}c > T_{N+(u,c)}(d)$, and thus $d \prec^T_{N+(u,c)} c$ as required.

For **Source-pos-resp**, suppose $s \in \mathrm{antisrc}_N(c)$, $t \in \mathrm{src}_N(c)$, $\mathrm{cl}_N(u) = \emptyset$ and $s \sqsubseteq^T_N t$. Then

$$
\begin{aligned}
T_{N+(u,c)}(t) - T_{N+(u,c)}(s) &= w_{N+(u,c)}(t) - w_{N+(u,c)}(s) \\
&= \underbrace{w_N(t) - w_N(s)}_{\geq 0} + \frac{\mathbb{1}[c \in \mathrm{cl}_N(t)]}{|\,\mathrm{cl}_N(t)|} - \underbrace{\frac{\mathbb{1}[c \in \mathrm{cl}_N(s)]}{|\,\mathrm{cl}_N(s)|}}_{=0} \\
&\geq \frac{1}{|\,\mathrm{cl}_N(t)|} \\
&> 0
\end{aligned}
$$

where we use the fact that $s \in \mathrm{antisrc}_N(c)$ means $c \notin \mathrm{cl}_N(s)$. Hence $s \sqsubset^T_{N+(u,c)} t$.

Finally, **Disjoint-independence** follows easily by noting that for disjoint networks N, N' and $s \in S_N$, $c \in C_N$, we have $\mathrm{cl}_{N \sqcup N'}(s) = \mathrm{cl}_N(s)$ and $\mathrm{src}_{N \sqcup N'}(c) = \mathrm{src}_N(c)$.

To see that **Source-coherence** does not hold, let N be the network shown in Fig. 2.10. One can easily check that $c \prec^T_N c'$ yet $s \simeq^T_N s'$.

Classical-independence cannot hold by the impossibility result Theorem 2.3.1 (1), since **Symmetry**, **Claim-coherence**, **Fresh-pos-resp** and **Source-pos-resp** have already been shown to hold. Similarly, the failure of **Conflict-coherence** and **Anti-coherence** follow from Theorem 2.3.2. $\square$

Sums. To simplify axiomatic analysis of Sums, we first show that T^{sums} is a fixed point of the update function U for Sums. In what follows, let $(\mathcal{D}, T^0, U)$ denote the recursive scheme corresponding to Sums from Definition 2.2.4. Recall that T^{sums} is defined as the limit of this recursive scheme. For simplicity

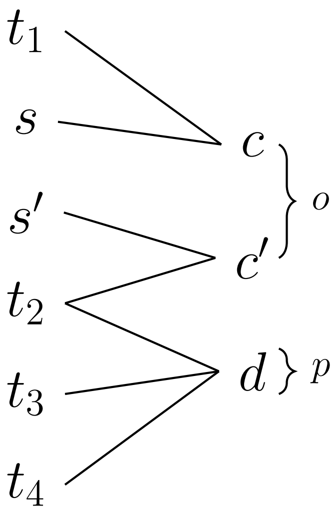


Figure 2.10: Counterexample for ***Source-coherence*** *for Weighted Agreement.*

we assume T^{sums} converges on all input networks.[9] We also write $T^n = U^n(T^0)$ for the n-th step of the iteration of Sums.

The following lemma helps to deal with the normalisation factors used in the update function for Sums.

Lemma 2.4.2. *Let $(x_n^i)_{n\in\mathbb{N}}$ be convergence sequences in $\mathbb{R}$, for $1 \le i \le k$. Then*

$$\lim_{n\to\infty} \max_i |x_n^i| = \max_i |\lim_{n\to\infty} x_n^i|.$$

Proof. Let $\varepsilon > 0$. Write $y^i = \lim_{n\to\infty} x_n^i$. For each i, hence $|x_n^i| \to |y^i|$ – since the absolute value function $\|\cdot\|$ is continuous – and so there is $n_i \in \mathbb{N}$ such that $||x_n^i| - |y^i|| < \varepsilon$ for all $n \ge n_i$. Take $m = \max_i n_i$. Let $n \ge m$. For any i, we have

$$|y^i| - \varepsilon < |x_n^i| < |y^i| + \varepsilon.$$

Thus

$$|x_n^i| < |y^i| + \varepsilon \le \max_j |y^j| + \varepsilon.$$

Since the maximum is achieved for some i, we get

$$\max_i |x_n^i| < \max_j |y^j| + \varepsilon. \tag{2.7}$$

Now, take j such that $\max_i |y^i| = |y^j|$. Then

$$\max_i |x_n^i| \ge |x_n^j| > |y^j| - \varepsilon = \max_i |y^i| - \varepsilon. \tag{2.8}$$

[9]While Pasternack and Roth (2010) do not consider convergence, Sums is an adaptation of the *Hubs and Authorities* algorithm, for which Kleinberg (1999) proves convergence: phrased in our terminology, he shows that the vector of source scores converge to a unit eigenvector of the matrix MM^{T} corresponding to the largest eigenvector (in absolute value), where M is the $|S| \times |C|$ matrix defined by $M_{sc} = \mathbb{1}[s \in \mathrm{src}_N(c)]$. Similarly, claim scores converge to a unit eigenvector of $M^{\mathrm{T}}M$.

Combining (2.7) and (2.8), we get

$$|\max_i |x_n^i| - \max_i |y^i|| < \varepsilon$$

as required. ☐

Lemma 2.4.3. *$T^{\text{sums}} \in \mathcal{D}$, and $U(T^{\text{sums}}) = T^{\text{sums}}$.*

Proof. Note that $T_N^n(z) \in [0,1]$ for all n and $z \in S \cup C$. Consequently $T_N^*(z) = \lim_{n\to\infty} T_N^n(z) \in [0,1]$, since $[0,1]$ is closed. Hence $T^{\text{sums}} \in \mathcal{D}$.

Take any network N. If N contains no reports – i.e. $R = \emptyset$, then $T^n \equiv 0$ for all $n > 1$. Hence $T_N^{\text{sums}} \equiv 0$ and $U(T^{\text{sums}})_N = T_N^{\text{sums}}$. Now suppose N contains at least one report (s_0, c_0). It is easily checked that in this case $T_N^n(s_0), T_N^n(c_0) > 0$ for all n. Consequently the maximums in the definition of α and β in Definition 2.2.4 are non-zero. For any $s \in S$, we therefore have

$$\begin{aligned} T_N^{\text{sums}}(s) &= \lim_{n\to\infty} T_N^n(s) \\ &= \lim_{n\to\infty} T_N^{n+1}(s) \\ &= \lim_{n\to\infty} \frac{\sum_{c\in \mathrm{cl}_N(s)} T_N^n(c)}{\max_{t\in S} \left|\sum_{c\in \mathrm{cl}_N(t)} T_N^n(c)\right|} \end{aligned} \tag{2.9}$$

We need to show that the denominator in (2.9) converges to a non-zero limit. By the normalisation step for claim scores, for each $n > 1$ there is a claim c_n with $|T_N^n(c_n)| = 1$. Since there are only finitely many claims, this implies we cannot have $T_N^{\text{sums}}(c) = 0$ for all c, so there is some c_1 with $T_N^{\text{sums}}(c_1) > 0$. Furthermore, $\mathrm{src}_n(c_1) \neq \emptyset$ (otherwise one can easily show $T_N^{\text{sums}}(c_1) = 0$). Likewise, there is some s_1 such that $T_N^{\text{sums}}(s_1) > 0$. Now using the fact that $T_N^n(c) \to T_N^{\text{sums}}(c)$ for each c and taking the limit of the sum, Lemma 2.4.2 gives

$$\lim_{n\to\infty} \max_{t\in S} \left| \sum_{c\in \mathrm{cl}_N(t)} T_N^n(c) \right| = \max_{t\in S} \left| \sum_{c\in \mathrm{cl}_N(t)} T_N^{\text{sums}}(c) \right| \geq T_N^{\text{sums}}(c_1) > 0.$$

Splitting the limit across the quotient in (2.9), we find

$$\begin{aligned} T_N^{\text{sums}}(s) &= \frac{\lim_{n\to\infty} \sum_{c\in \mathrm{cl}_N(s)} T_N^n(c)}{\lim_{n\to\infty} \max_{t\in S} \left|\sum_{c\in \mathrm{cl}_N(t)} T_N^n(c)\right|} \\ &= \frac{\sum_{c\in \mathrm{cl}_N(s)} T_N^{\text{sums}}(c)}{\max_{t\in S} \left|\sum_{c\in \mathrm{cl}_N(t)} T_N^{\text{sums}}(c)\right|} \\ &= U(T^{\text{sums}})_N(s) \end{aligned}$$

as required. One can show $T_N^{\text{sums}}(c) = U(T^{\text{sums}})_N(c)$ for any claim c by a near-identical argument, and thus $U(T^{\text{sums}})_N = T_N^{\text{sums}}$. Since N was arbitrary this shows $U(T^{\text{sums}}) = T^{\text{sums}}$, and the proof is complete. ☐

Corollary 2.4.2. T^{sums} *is weightable.*

Proof. We define a weighting w as follows. If N contains no reports, set $w_N \equiv 0$. Otherwise, set

$$w_N(s) = \frac{T_N^{\text{sums}}(s)}{\max_{c \in C} \left| \sum_{t \in \mathrm{src}_N(c)} T_N^{\text{sums}}(t) \right|}. \tag{2.10}$$

We need to show $T^{\text{sums}} \sim T^w$, i.e. that T^{sums} and T^w give the same rankings on all networks N. If N contains no reports then both T_N^{sums} and T^w are zero, and therefore output the same rankings. Suppose N contains at least one report. Since we just divide by a constant in (2.10), $s \sqsubseteq_N^{T^{\text{sums}}} s'$ iff $s \sqsubseteq_N^{T^w} s'$ for all sources s and s'. Using the fact that $T^{\text{sums}} = U(T^{\text{sums}})$ from Lemma 2.4.3, it is easily seen that $T_N^{\text{sums}}(c) = \sum_{s \in \mathrm{src}_N(c)} w_N(s) = T_N^w(c)$. Hence T_N^{sums} and T_N^w give exactly the same scores for claims, and in particular the rankings also coincide. □

We come to the axioms satisfied by Sums. While it satisfies both **Claim-coherence** and **Source-coherence**, it is notable that Sums fails both monotonicity properties and **Disjoint-independence**. In some sense these problems are caused by the normalisation step, where source and claim scores are divided by their respective maximums. We present a modified version of Sums without these deficiencies in Section 2.4.1.

Theorem 2.4.3. *Sums satisfies* ***Claim-coherence****,* ***Source-coherence****,* ***Symmetry****,* ***Marginal-trustworthiness****,* ***Trust-based-monotonicity****. It does not satisfy* ***Fresh-pos-resp****,* ***Source-pos-resp****,* ***Classical-independence****,* ***Disjoint-independence****,* ***Conflict-coherence*** *or* ***Anti-coherence****.*

Proof. **Claim-coherence**, **Marginal-trustworthiness** and **Trust-based-monotonicity** follow directly from Corollaries 2.4.1 and 2.4.2. For **Source-coherence**, let N be a network and suppose $\mathrm{cl}_N(s)$ strictly precedes $\mathrm{cl}_N(s')$ with respect to $\preceq_N^{T^{\text{sums}}}$. Then by Proposition 2.3.1, there is a bijection $f : \mathrm{cl}_N(s) \to \mathrm{cl}_N(s')$ such that $T_N^{\text{sums}}(c) \leq T_N^{\text{sums}}(f(c))$ for all $c \in \mathrm{cl}_N(s)$, and there is some c_0 with $T_N^{\text{sums}}(c_0) < T_N^{\text{sums}}(f(c_0))$. It follows that N must contain at least one report, since otherwise no strict inequalities hold. For any source t, Lemma 2.4.3 implies

$$T_N^{\text{sums}}(t) = \alpha \sum_{c \in \mathrm{cl}_N(t)} T_N^{\text{sums}}(c),$$

where $\alpha = 1 / \max_{t' \in S} |\sum_{c \in \mathrm{cl}_N(t')} T_N^{\text{sums}}(c)| > 0$ is a constant. Using the fact

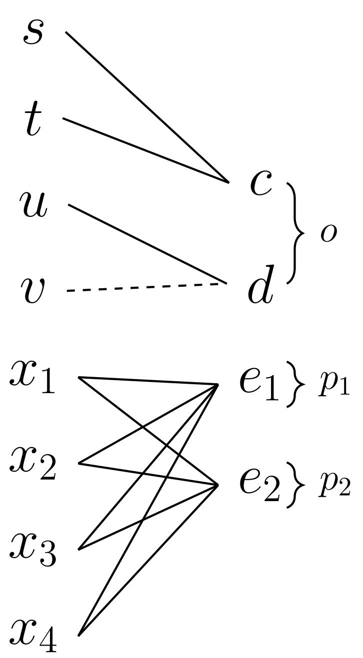


Figure 2.11: Networks used as counterexamples for Sums axiom failures.

that f maps bijectively from $\mathrm{cl}_N(s)$ to $\mathrm{cl}_N(s')$, we get

$$\begin{aligned} T^{\mathrm{sums}}_N(s) - T^{\mathrm{sums}}_N(s') &= \alpha \left(\sum_{c \in \mathrm{cl}_N(s)} T^{\mathrm{sums}}_N(c) - \sum_{c' \in \mathrm{cl}_N(s')} T^{\mathrm{sums}}_N(c') \right) \\ &= \alpha \left(\sum_{c \in \mathrm{cl}_N(s)} T^{\mathrm{sums}}_N(c) - \sum_{c \in \mathrm{cl}_N(s)} T^{\mathrm{sums}}_N(f(c)) \right) \\ &= \alpha \sum_{c \in \mathrm{cl}_N(s)} (T^{\mathrm{sums}}_N(c) - T^{\mathrm{sums}}_N(f(c))). \end{aligned}$$

By assumption $T^{\mathrm{sums}}_N(c) - T^{\mathrm{sums}}_N(f(c)) \leq 0$ for each c, and the inequality is strict for $c = c_0$. Hence $T^{\mathrm{sums}}_N(s) < T^{\mathrm{sums}}_N(s')$, and $s \sqsubset^{T^{\mathrm{sums}}}_N s'$ as required.

Finally, **Symmetry** can be shown in similar way to Weighted Agreement, since Sums is defined only in terms of src_N and cl_N.

For the negative axioms, we refer to networks shown in Fig. 2.11. For **Fresh-pos-resp** and **Source-pos-resp**, let N_0 denote the network without the dashed report (v, d), so that $N_0 + (v, d)$ is the full network. It can be shown that the rankings are the same under Sums in both networks, with $s \simeq t \simeq u \simeq v \sqsubset x_1 \simeq x_2 \simeq x_3 \simeq x_4$ and $c \approx d \prec e_1 \approx e_2$. This violates **Fresh-pos-resp**, since $c \preceq^{T^{\mathrm{sums}}}_{N_0} d$ but $c \not\prec^{T^{\mathrm{sums}}}_{N_0+(v,d)} d$. It also violates **Source-pos-resp**, since $s \in \mathrm{antisrc}_{N_0}(d)$, $u \in \mathrm{src}_n(d)$ and $s \sqsubseteq^{T^{\mathrm{sums}}}_{N_0} u$, but $s \not\sqsubset^{T^{\mathrm{sums}}}_{N_0+(v,d)} u$.

For **Classical-independence** and **Disjoint-independence**, let N_1 and N_2 denote the upper and lower components of the network in Fig. 2.11, excluding the dashed report (v, d). Then $N_0 = N_1 \sqcup N_2$. Hence $c \approx^{T^{\mathrm{sums}}}_{N_1 \sqcup N_2} d$. However, it is straightforward to check that in the network N_1 alone we have $d \prec_{N_1}^{T^{\text{sums}}} c$; this violates **Disjoint-independence**. Taking N_0' to be the network obtained from N_0 by removing all reports from $x_1, \ldots, x_4$, we have $d \prec_{N_0'}^{T^{\text{sums}}} c$ and $c \approx_{N_0}^{T^{\text{sums}}} d$. Since c and d have the same sources in both networks, this violates **Classical-independence**.

Finally, for **Conflict-coherence** and **Anti-coherence** we can reuse the network N from Fig. 2.7 (including the dashed report). Applying Sums to this network, we have $T_N^{\text{sums}}(s) = T_N^{\text{sums}}(t) = 0$, $T_N^{\text{sums}}(u) = \sqrt{3} - 1 \approx 0.7321$, $T_N^{\text{sums}}(s') = T_N^{\text{sums}}(t') = 1$ and $T_N^{\text{sums}}(c) = T_N^{\text{sums}}(d) = T_N^{\text{sums}}(e) = 0$, $T_N^{\text{sums}}(f) = 1$, $T_N^{\text{sums}}(c') = T_N^{\text{sums}}(d') = \frac{1}{2}(\sqrt{3} - 1) \approx 0.3660$, yielding rankings $s \simeq t \sqsubset u \sqsubset s' \simeq t'$ and $c \approx d \approx e \prec c' \approx d' \prec f$. This ranking violates **Conflict-coherence** since $\text{conflict}_N(c) = \{d\}$ strictly precedes $\text{conflict}_N(c') = \{d'\}$ but $c' \not\prec_N^{T^{\text{sums}}} c$. It also violates **Anti-coherence**, since $\text{antisrc}_N(c) = \{t\}$ strictly precedes $\text{antisrc}_N(c') = \{t'\}$ but $c' \not\prec_N^{T^{\text{sums}}} c$. $\square$

The key to the counterexamples derived from Fig. 2.11 in the above proof lies in the lower disjoint component, which takes the form of a *connected* bipartite graph. That is, each source x_i reports each claim e_j in the component. Moreover, sources elsewhere in the network claim fewer facts than the x_i, and claims elsewhere are reported by fewer sources than the e_j.

Since Sums assigns scores by a simple sum, this results in the scores for the x_i and e_j dominating those of the other sources and claims. The normalisation step (i.e. the factors α and β in Definition 2.2.4) then divides these scores by the (comparatively large) maximum. As the next result shows, under certain conditions this causes scores to decrease *exponentially* and become 0 in the limit. In particular, we can generate pathological examples such as Fig. 2.11 where a whole component receives scores of 0, which leads to failure of **Disjoint-independence** and the monotonicity axioms.

Proposition 2.4.1. *Let N be a network. Suppose there are non-empty sets $X \subseteq S$, $Y \subseteq C$ such that*

1. $\text{cl}_N(x) = Y$ *for each* $x \in X$*;*
2. $\text{src}_N(y) = X$ *for each* $y \in Y$*;*
3. $|\text{cl}_N(s)| \leq \frac{|Y|}{2}$ *for each* $s \in S \setminus X$*; and*
4. $|\text{src}_N(c)| \leq \frac{|X|}{2}$ *for each* $f \in C \setminus Y$*.*

Then, with T^n denoting the n-th step in the iteration of Sums, for all $n > 0$ we have

$$T_N^n(s) \leq \frac{1}{2^{n-1}} \qquad (s \in S \setminus X),$$

$$T_N^n(c) \leq \frac{1}{2^{n-1}} \qquad (c \in C \setminus Y),$$

$$T_N^n(x) = 1, \qquad (x \in X)$$

$$T_N^n(y) = 1. \qquad (y \in Y)$$

In particular, $T_N^{\text{sums}}(s) = T_N^{\text{sums}}(c) = 0$ *for all* $s \in S \setminus X$ *and* $c \in C \setminus Y$*.*

The proof follows by induction on n.

2.4.1 Modifying Sums

While Sums satisfies some desirable axioms, the failure of **Disjoint-independence** and the monotonicity axioms is problematic. We saw through Proposition 2.4.1 that this is in some sense caused by the normalisation step, where source and claim scores are divided by the "global" maximum scores across the network.

A seemingly natural fix for **Disjoint-independence** is to therefore use different normalisation factors α and β for each disjoint component. However, this does not escape the negative consequences of Proposition 2.4.1. Indeed, if one modifies the network(s) in Fig. 2.11 so that claim e_1 is associated with object o instead of p_1, the network no longer has two disjoint components. Consequently, the "per-component Sums" operator gives the same result as Sums itself, and in particular the counterexamples for **Fresh-pos-resp** and **Source-pos-resp** still hold. Perhaps even worse, one can show that **Claim-coherence** fails for this modified operator. We consider loss of **Claim-coherence** too high a price to pay for **Disjoint-independence**.

Instead, let us take a step back and consider if normalisation is truly necessary. On the one hand, without normalisation the source and claim scores are unbounded and therefore do not converge. On the other, we are not interested in the numeric scores for their own sake, but rather for the *rankings* that they induce. It may be possible that whilst the scores diverge without normalisation, the induced rankings *do* converge to a fixed one, which we may take as the "ordinal limit". This is in fact the case. We call this new operator *UnboundedSums*

Definition 2.4.1. UnboundedSums *is the recursive scheme* $(\mathcal{D}, T^0, U)$*, where* $\mathcal{D}$ *is the set of all operators with scores in* $[0, 1]$*,*

$$T_N^0(s) = 1,$$
$$T_N^0(c) = |\operatorname{src}_N(c)|$$

and $U(T) = T'$*, where*

$$T'_N(s) = \sum_{c \in \operatorname{cl}_N(s)} T_N(c),$$
$$T'_N(c) = \sum_{s \in \operatorname{src}_N(c)} T'_N(s).$$

Note that the update function U is almost identical to that of Sums from Definition 2.2.4, expect that it does not include the normalisation factors α

and β. Also note that to simplify the proof of ordinal convergence, we use a different initial operator T^0 compared to Sums. In what follows, let T^n denote the n-th step in the iteration of UnboundedSums.

Theorem 2.4.4. *UnboundedSums is ordinally convergent in the following sense: for every network N there is $m \in \mathbb{N}$ such that for all $n \geq m$, $s, s' \in S$ and $c, c' \in C$,*

$$\begin{aligned} T^n_N(s) \leq T^n_N(s') &\iff T^m_N(s) \leq T^m_N(s'), \\ T^n_N(c) \leq T^n_N(c') &\iff T^m_N(c) \leq T^m_N(c'). \end{aligned}$$

That is, the rankings induced by T^n_N are constant for $n \geq m$.

The proof – which can be found in Appendix A.1 – expresses the update function of UnboundedSums in terms of matrix multiplication, and uses techniques and results from linear algebra. In light of Theorem 2.4.4, we may define an operator $T^{\text{u-sums}}$ by setting $T^{\text{u-sums}}_N(z) = T^m_N(z)$, where m depends on N and is taken sufficiently large. It follows that $U(T^{\text{u-sums}}) \sim T^{\text{u-sums}}$.

With the normalisation problems aside, UnboundedSums provides an example of a principled operator satisfying many of our core axioms, including **Source-coherence**, **Claim-coherence** and **Disjoint-independence**. In particular, we avoid the undesirable behaviour of Sums in Fig. 2.11; whereas Sums trivialises the ranking of sources and claims in the upper component, UnboundedSums allows meaningful comparison (e.g. we have $d \prec c$). We conjecture that UnboundedSums also satisfies the monotonicity properties **Fresh-pos-resp** and **Source-pos-resp**, but this remains to be proven.[10]

Theorem 2.4.5. *UnboundedSums satisfies* ***Source-coherence****,* ***Claim-coherence****,* ***Symmetry****,* ***Marginal-trustworthiness****,* ***Trust-based-monotonicity*** *and* ***Disjoint-independence****. It does not satisfy* ***Classical-independence****,* ***Conflict-coherence*** *or* ***Anti-coherence****.*

Proof (sketch). Setting $w_N(s) = U(T^{\text{u-sums}})_N(s) = \sum_{c \in \text{cl}_n(s)} T^{\text{u-sums}}_n(c)$, one can show $T^{\text{u-sums}} \sim T^w$, and thus UnboundedSums is weightable. By Corollary 2.4.1, **Claim-coherence**, **Marginal-trustworthiness** and **Trust-based-monotonicity** hold.

For **Source-coherence**: if $\text{cl}_N(s)$ strictly precedes $\text{cl}_N(s')$ with respect to $\preceq^{T^{\text{u-sums}}}_N$, then taking the same approach as in the proof of **Source-coherence** for Sums, we have

$$\sum_{c \in \text{cl}_N(s)} T^{\text{u-sums}}_N(c) < \sum_{c' \in \text{cl}_N(s')} T^{\text{u-sums}}_N(c')$$

and thus $U(T^{\text{u-sums}})_N(s) < U(T^{\text{u-sums}})_N(s')$. Since $T^{\text{u-sums}} \sim U(T^{\text{u-sums}})$, we get $s \sqsubset^{T^{\text{u-sums}}}_N s'$ as required.

[10]We have experimentally verified that UnboundedSums satisfies all the specific instances of **Fresh-pos-resp** with the starting network N as in Fig. 2.2.

Symmetry holds by an argument similar to the one employed for Weighted Agreement.

For **Disjoint-independence**, suppose N and N' are disjoint. One can show (e.g. by induction) that $T^n_N(z) = T^n_{N\sqcup N'}(z)$ for all $z \in S \cup C$ and $n \in \mathbb{N}$. Taking n sufficiently large, T^n_N and $T^n_{N\sqcup N'}$ yield the same rankings on S and C as $T^{\text{u-sums}}_N$ and $T^{\text{u-sums}}_{N\sqcup N'}$ respectively. Consequently, for any $s, t \in S$,

$$\begin{aligned} s \sqsubseteq^{T^{\text{u-sums}}}_N t &\iff T^n_N(s) \leq T^n_N(t) \\ &\iff T^n_{N\sqcup N'}(s) \leq T^n_{N\sqcup N'}(t) \\ &\iff s \sqsubseteq^{T^{\text{u-sums}}}_{N\sqcup N'} t. \end{aligned}$$

Similarly, $c \preceq^{T^{\text{u-sums}}}_N d$ iff $c \preceq^{T^{\text{u-sums}}}_{N\sqcup N'} d$ for $c, d \in C$, and **Disjoint-independence** is shown.

To see **Classical-independence** does not hold, consider the network N from Fig. 2.2 and the network N' obtained by removing reports from sources u and v. Then one can show $c \prec^{T^{\text{u-sums}}}_N d$ but $c \approx^{T^{\text{u-sums}}}_{N'} d$, so **Classical-independence** fails. Finally, the same counterexamples for **Conflict-coherence** and **Anti-coherence** for Sums provide counterexamples for UnboundedSums. □

2.5 Related Work

Ranking systems. Altman and Tennenholtz (2008) initiated axiomatic study of ranking systems. First we discuss their framework in relation to ours, and then turn to their axioms. In their framework, a *ranking system* F maps any (finite) directed graph $G = (V, E)$ to a total preorder $\leq^F_G$ on the vertex set V. In their view this is a variation of the classical social choice setting, in which the set of voters and alternatives coincide. Nodes $v \in V$ "vote" on their peers in V by a form of approval voting (Laslier and Sanver 2010): an edge $v \to u$ is interpret as a vote for u from v. A ranking system then outputs a ranking of V, following the general intuition that the more "votes" v receives (i.e. the more incoming edges), the higher v should rank. As with the ranking of claims in truth discovery, this does not necessarily mean ranking nodes simply by the *number* of votes received, since the *quality* of the voters should also be taken in account. For example, a ranking system may prioritise nodes which receive few votes from highly ranked nodes over those with many votes from lower ranked nodes.

Note that our truth discovery networks N can be viewed as directed graphs on the vertex set $S \cup C \cup O$; this is the presentation we have used in the figures throughout this chapter. However, naively applying a ranking system to such graphs directly makes little sense: sources never receive any "votes", and the resulting ranking includes objects, which do not need to be ranked in our setting. Perhaps a more sensible approach is to consider the bipartite graph

$G_N = (V_N, E_N)$ associated with a network N, where

$$V_N = S \cup C, \qquad E_N = \bigcup_{(s,c)\in R} \{(s,c),(c,s)\}.$$

That is, we take the edges from sources to claims together with the reversal of such edges. The edges in G_N have some intuitive interpretation: a source votes for the claims which it believes are true, and a claim votes for the sources who vouch for it. Any ranking system F thus gives rise to a truth discovery operator, where $s \sqsubseteq_N^T t$ iff $s \leq_{G_N}^F t$, and similar for claims.

However, some characteristic aspects of the truth discovery problem are lost in this translation to ranking systems. Notably, the objects play no role at all in G_N. Sources and claims are also treated symmetrically, where they perhaps should not be. For example, a claim c receiving more claims than d is beneficial for c, all else being equal, but a source s reporting more claims than t does not tell us anything about the relative trustworthiness of s and t.

While other choices of G_N may be possible to alleviate some of these problems, we believe the truth discovery is sufficiently specialised beyond general graph ranking so that a bespoke modelling is required to capture its nuances appropriately. Our framework provides this novel contribution.

In (Altman and Tennenholtz 2008), Altman and Tennenholtz also introduce axioms for ranking systems. Their first set of axioms deal with the transitive effects of voting when the alternatives are the voters themselves. As mentioned in Section 2.3.1, these axioms provided the inspiration for our coherence axioms. The core idea is that if the predecessors of a node v are weaker than those of u, then v should be ranked below u. If v additionally has *more* predecessors, v should rank *strictly* below. Coherence applies this idea both in the direction of sources-to-claims (**Claim-coherence**) and from claims-to-sources (**Source-coherence**). A notable difference is that we only consider the case where the number of sources for two claims (or the number of claims, for two sources) is the same. For example, a source reporting more claims does not give it the strict boost Transitivity would dictate. Under the mapping $N \mapsto G_N$ described above, any ranking system satisfying Transitivity induces a truth discovery operator which satisfies both **Source-coherence** and **Claim-coherence**.

The other axiom of Altman and Tennenholtz (2008) is the independence axiom *RIIA* (ranked independence of irrelevant alternatives), which adapts the classical IIA axiom from social choice theory to the ranking system setting, although in a different manner to our independence axioms of Section 2.3.4. We describe the axiom in rough terms, deferring to the paper for the technical details. Suppose the relative ranking of u_1's predecessors compared to u_2's predecessors is the same as that of v_1's compared to v_2's. Then RIIA requires $u_1 \leq u_2$ iff $v_1 \leq v_2$. Informally, "the relative ranking of two agents must only depend on the pairwise comparison of the ranks of their predecessors" (Altman and Tennenholtz 2008). While we do not have an analogous axiom, the idea can be adapted to truth discovery networks. Intuitively, such an axiom would state that the ranking of two claims depends only on comparisons between their corresponding sources (and similar for the ranking of sources).

However, the main result of Altman and Tennenholtz is an impossibility: Transitivity is incompatible with RIIA. Moreover, the result remains true even when restricting to bipartite graphs, such as G_N described above. Accordingly, we can expect a similar impossibility result to hold in the truth discovery setting between the coherence axioms and any analogue of RIIA.

PageRank. PageRank (Page et al. 1999) is a well-known algorithm for ranking web pages based on the hyperlink structure of the web, viewed as a directed graph. It has also been studied through the lens of social choice and characterised axiomatically (Altman and Tennenholtz 2005; Wąs and Skibski 2020).[11]

Altman and Tennenholtz (2005) propose several *invariance axioms*, each of which requires that the ranking of pages is not affected by a certain transformation of the web graph. For example, the axiom *Self Edge* says that adding a self loop from a page a to itself does not change the relative ranking of other pages, and results in a strictly positive boost for a (c.f. our monotonicity axioms). However, if we identify a truth discovery network N with the graph G_N as described above, most of the transformations involved do not respect the bipartite, symmetric structure of G_N. That is, the transformed graph does not correspond to any $G_{N'}$, for a network N'. Consequently, the PageRank axioms have no truth discovery counterpart in our setting. The only exception is *Isomorphism*, where the transformation in question is graph isomorphism; this axiom is analogous to our **Symmetry** axiom. However, since PageRank is similar to the *Hubs and Authorities* (Kleinberg 1999) algorithm on which Sums is based – which also uses the hyperlink structure of the web to rank pages – we expect there may be additional axioms which can be expressed both for general graphs *and* truth discovery networks, satisfied by PageRank and Sums. We leave the task of finding such axioms to future work.

2.6 Conclusion

Summary. In this chapter we formalised a mathematical framework for truth discovery. While a number of simplifying assumptions were made compared to the mainstream truth discovery literature, we are able to express several algorithms in the framework. This provided the setting for the axiomatic method of social choice to be applied. To our knowledge, this is the first such axiomatic treatment in this context.

[11] Wąs and Skibski (2020) axiomatise the *numerical scores* of PageRank, whereas Altman and Tennenholtz (2005) axiomatise the resulting ranking. Moreover, Wąs and Skibski point out that Altman and Tennenholtz in fact only consider a simplified version of PageRank called *Katz prestige*, defined only for strongly connected graphs.

It was possible to adapt many axioms from social choice theory and related areas. In particular, the *Transitivity* axiom studied in the context of ranking systems (Tennenholtz 2004; Altman and Tennenholtz 2008) took on new life in the form of the coherence axioms, which we consider essential for truth discovery operators. We proceeded to highlight the differences between voting theory and truth discovery via an impossibility result involving a classical independence axiom, and used this axiom to characterise Voting. Another impossibility result – Theorem 2.3.2 – established the tension between methods which rank claims on the basis of their sources, and those which rank on the basis of antisources.

On the practical side, we analysed the existing method Sums, and found that, surprisingly, it fails **Disjoint-independence**. This is a serious issue for Sums which has not been discussed in the literature to date, and its discovery here highlights the benefits of the axiomatic method. To resolve this, we suggested a modification to Sums – which we call UnboundedSums – for which **Disjoint-independence** *is* satisfied.

Limitations and future work. While UnboundedSums resolves axiomatic problems with Sums, it may introduce computational difficulties (since the numeric scores involved grow without bound). Moreover, its status with respect to the monotonicity axioms remains unknown, although we suspect that the axioms do hold. We leave further investigation of UnboundedSums to future work.

A restriction of our analysis is that only one algorithm from the literature was studied in detail. Further axiomatic analysis of algorithms provides a deeper understanding of how algorithms operate on an intuitive level, but is made difficult by the complexity of the state-of-the-art truth discovery methods. New techniques for establishing the satisfaction (or otherwise) of axioms would be helpful in this regard.

There is also scope for extensions to our model of truth discovery in the framework itself. For example, we make the somewhat simplistic assumption that there are only finitely many possible facts for sources to claim. This effectively means we can only consider *categorical values*; modelling an object whose domain is the set of real numbers, for example, is not straightforward in our framework.

Next, our model does not account for any associations or constraints between objects, whereas in reality the belief in a fact for one object may strengthen or weaken our belief in other facts for related objects. These types of constraints or correlations have been studied both on the theoretical side (e.g. in judgement aggregation) and practical side in truth discovery (Yang, Bai, and Liu 2019).

The axioms can also be further refined to relax some of the simplifying assumptions we make regarding source attitudes; e.g. that they do not collude or attempt to manipulate.

Finally, it may be argued that truth discovery as formulated in this chapter risks simply to find *consensus* among sources, rather than the *truth*. In particular, we do not put forward any model for possible states of the world, nor of how sources produce their reports (c.f. (Meir et al. 2019)). Without such ingredients one cannot make precise what it means to find the truth. In a sense this is by design: our goal in applying the axiomatic approach is to find general principles which should hold for truth discovery methods under *any* notion of truth-tracking. However, truth-tracking will be addressed in the second half of the thesis – in Chapter 6 – although using a logic-based framework in the style of belief revision as opposed to truth discovery.

Outlook. In the following chapter we continue with a social-choice-based approach, and tackle the problem of *bipartite tournament ranking*. This is related to truth discovery – and we again consider ranking-based operators and axioms governing them – but is a separate problem in its own right. But in fact, Chapters 5 and 6 are more closely connected to truth discovery. We consider a logic-based belief change problem, in which an operator takes as input reports from several sources of unknown trustworthiness, and produces a conjecture concerning the true facts of the world and the trustworthiness of the sources themselves. A major point of difference as compared to truth discovery, however, is that we impose strict semantics on “trustworthiness”, rooted in *expertise*. That is, a source is considered trustworthy if they are believed to be an expert in the relevant domain.

— 3 —

Bipartite Tournaments

A tournament consists of a finite set of players equipped with a *beating relation* describing pairwise comparisons between each pair of players. Determining a ranking of the players in a tournament has applications in *voting* (Brandt, Brill, et al. 2016), where players represent alternatives and x beats y if a majority of voters prefer x over y, *paired comparisons analysis* (González-Díaz, Hendrickx, and Lohmann 2014), where players represent products and the beating relation expresses the preferences of a consumer, *search engines* (Slutzki and Volij 2006), *sports* (Bozóki, Csató, and Temesi 2016) and other domains.

In this chapter we introduce *bipartite tournaments*, which consist of two disjoint sets of players A and B such that comparisons only take place between players from opposite sets. We consider ranking methods which produce two rankings for each tournament – one for each side of the bipartition. Such tournaments model situations in which two different kinds of entity compete *indirectly* via matches against entities of the opposite kind.

The notion of competition may be abstract, which allows the model to be applied in a variety of settings. However, the principal motivation in the context of this thesis is the ranking of information sources by *expertise* or *trustworthiness*, as expressed by the following example.

Example 3.0.1. *Consider a truth discovery setting in which information sources* $\{a_1, \ldots, a_m\}$ *provide possible values to a number of objects. Among these objects, the "ground truth" values are known for a subset* $\{b_1, \ldots, b_n\}$*, and thus for any pair* (a, b) *it is known whether a provided the correct value or not. A natural question arises: how can the sources be ranked based on their trustworthiness?*

The straightforward approach of simply counting the number of correct values fails when the objects vary in difficulty. *Indeed, it may be preferable to reward sources for correct values on difficult objects, or to penalise them for failing on easy objects. Furthermore, the notion of difficulty is not intrinsic to an object, but depends on how easily sources are able to determine its correct value.*[1]

The setting of bipartite tournament ranking addresses these issues. Indeed, the two kinds of "players" are the sources and objects; a source "beats" a *object by providing the correct value, and otherwise the object beats the source. While we wish to compare sources based on their trustworthiness and objects based on their difficulty, there are no direct source-to-source or object-to-objects comparisons available: the ranking must be constructed on the basis of the indirect patterns of correctness between the set of sources and objects.*

Note that this is related to but not the same as the ranking problem of truth discovery itself, as studied in Chapter 2, since it does not concern finding the true values associated with objects. The setting of Example 3.0.1 is clearly relevant for *semi-supervised truth discovery*, in which a subset of ground truth is available (Yin and Tan 2011; Rekatsinas et al. 2017). However, even when no such ground truth is available, many recursive operators (including those described in Section 2.2.2) iteratively update source trust scores on the basis of current *estimates* of the true values for objects. One could therefore consider a bipartite tournament at each stage of the iteration. This may lead to *difficulty-aware* truth discovery methods (c.f. Galland et al. (2010)).

A related example is education (Jiao, Ravi, and Gatterbauer 2017), where A represents students, B exam questions, and student a "beats" question b by answering it correctly. The ranking of students then reflects their proficiency, and the ranking of questions reflects their difficulty. This may be particularly useful in the context of automated grading of crowdsourced questions provided by students themselves, which may vary in their difficulty (see for example the PeerWise system (Denny et al. 2008)).

Other application domains include the evaluation of generative models in machine learning (Olsson et al. 2018) (where A represents generators and B discriminators) and solo sports contests (e.g. where A represents golfers and B golf courses). In the remainder of the chapter we take the abstract view in which A and B are simply "players", without any fixed interpretation. However, we keep Example 3.0.1 in mind as our motivating example.

In principle, bipartite tournaments are a special case of *generalised* tournaments (González-Díaz, Hendrickx, and Lohmann 2014; Slutzki and Volij 2005; Csató 2019), which allow intensities of victories and losses beyond a binary win or loss (thus permitting draws or multiple comparisons), and do not require that every player is compared to all others. However, many existing ranking methods in the literature do not apply to bipartite tournaments due to the violation of an *irreducibility* requirement, which requires that the tournament graph is strongly connected. In any case, bipartite tournament ranking presents a unique problem – since we aim to rank players with only indirect information available – which we believe is worthy of study in its own right.

In this work we focus particularly on ranking via *chain graphs* and *chain editing*. A chain graph is a bipartite graph in which the neighbourhoods of vertices on one side form a chain with respect to set inclusion. A (bipartite)

[1] This is reminiscent of the mutual dependence between trustworthiness of sources and belief in claims in truth discovery.

tournament of this form represents an "ideal" situation in which the capabilities of the players are perfectly nested: weaker players defeat a subset of the opponents that stronger players defeat. In this case a natural ranking can be formed according to the set of opponents defeated by each player. These rankings respect the tournament results in an intuitive sense: if a player a defeats b and b' ranks worse than b, then a must defeat b' also. Unfortunately, this perfect nesting may not hold in reality: a weak player may win a difficult match by coincidence, and a strong player may lose a match by accident. With this in mind, Jiao, Ravi, and Gatterbauer (2017) suggested an appealing ranking method for bipartite tournaments: apply *chain editing* to the input tournament – i.e. find the minimum number of edge changes required to form a chain graph – and output the corresponding rankings. Whilst their work focused on algorithms for chain editing and its variants, we look to study the properties of the ranking method itself through the lens of computational social choice.

Contributions. Our primary contribution is the introduction of a class of ranking mechanisms for bipartite tournaments defined by chain editing. We also provide a new probabilistic characterisation of chain editing via maximum likelihood estimation. To our knowledge this is the first in-depth study of chain editing as a ranking mechanism. Secondly, we introduce a new class of "chain-definable" mechanisms by relaxing the minimisation constraint of chain editing in order to obtain tractable algorithms and to resolve the failure of an important anonymity axiom. We present a concrete example of such an algorithm, and characterise it axiomatically.

This chapter is an extension of Singleton and Booth (2021), with new results presented in Section 3.5.4.

3.1 Preliminaries

In this section we define our framework for bipartite tournaments, introduce chain graphs and discuss the link between them.

3.1.1 Bipartite Tournaments

Following the literature on generalised tournaments (González-Díaz, Hendrickx, and Lohmann 2014; Slutzki and Volij 2005; Csató 2019), we represent a tournament as a matrix, whose entries represent the results of matches between participants. In what follows, $[n]$ denotes the set $\{1, \dots, n\}$ whenever $n \in \mathbb{N}$.

Definition 3.1.1. *A* bipartite tournament *– hereafter simply a* tournament *– is a triple (A, B, K), where $A = [m]$ and $B = [n]$ for some $m, n \in \mathbb{N}$, and K is an $m \times n$ matrix with $K_{ab} \in \{0, 1\}$ for all $(a, b) \in A \times B$. The set of all tournaments will be denoted by $\mathcal{K}$.*

Here A and B represent the two sets of players in the tournament.[2] An entry K_{ab} gives the result of the match between $a \in A$ and $b \in B$: it is 1 if a defeats b and 0 otherwise. Note that we do not allow for the possibility of draws, and every $a \in A$ faces every $b \in B$. When there is no ambiguity we denote a tournament simply by K, with the understanding that $A = [\text{rows}(K)]$ and $B = [\text{columns}(K)]$.

The *neighbourhood* of a player $a \in A$ in K is the set $K(a) = \{b \in B \mid K_{ab} = 1\} \subseteq B$, i.e. the set of players which a defeats. The neighbourhood of $b \in B$ is the set $K^{-1}(b) = \{a \in A \mid K_{ab} = 1\} \subseteq A$, i.e. the set of players defeating b.

Given a tournament K, our goal is to place a ranking on each of A and B. We define a tournament ranking operator for this purpose.

Definition 3.1.2. *A* tournament ranking operator T *assigns each tournament* K *a pair* $T(K) = (\preceq^T_K, \sqsubseteq^T_K)$ *of total preorders on* A *and* B *respectively.*

For $a, a' \in A$, we interpret $a \preceq^T_K a'$ to mean that a' is ranked *at least as strong* as a in the tournament K, according to the operator T (similarly, $b \sqsubseteq^T_K b'$ means b' is ranked at least as strong as b). The strict and symmetric parts of $\preceq^T_K$ are denoted by $\prec^T_K$ and $\approx^T_K$.

As a simple example, consider T_{count}, where $a \preceq^{T_{\mathsf{count}}}_K a'$ iff $|K(a)| \leq |K(a')|$ and $b \sqsubseteq^{T_{\mathsf{count}}}_K b'$ iff $|K^{-1}(b)| \geq |K^{-1}(b')|$. This operator simply ranks players by number of victories. It is a bipartite version of the *points system* introduced by Rubinstein (1980), and generalises *Copeland's rule* (Brandt, Brill, et al. 2016).

3.1.2 Chain Graphs

Each bipartite tournament K naturally corresponds to a bipartite graph G_K, with vertices $A \sqcup B$ and an edge between a and b whenever $K_{ab} = 1$.[3] The task of ranking a tournament admits a particularly simple solution if this graph happens to be a *chain graph*.

Definition 3.1.3 (Yannakakis (1981))**.** *A bipartite graph* $G = (U, V, E)$ *is a* chain graph *if there is an ordering* $U = \{u_1, \ldots, u_k\}$ *of* U *such that* $N(u_1) \subseteq \cdots \subseteq N(u_k)$*, where* $N(u_i) = \{v \in V \mid (u_i, v) \in E\}$ *is the neighbourhood of* u_i *in* G*.*

In other words, a chain graph is a bipartite graph where the neighbourhoods of the vertices on one side can be ordered so as to form a chain with respect

[2] Note that A and B are not disjoint as sets: 1 is always contained in both A and B, for instance. This poses no real problem, however, since we view the number 1 merely a *label* for a player. It will always be clear from context whether a given integer should be taken as a label for a player on the A side or the B side.

[3] $A \sqcup B$ is the *disjoint union* of A and B, which we define as $\{(a, 0) \mid a \in A\} \cup \{(b, 1) \mid b \in B\}$.

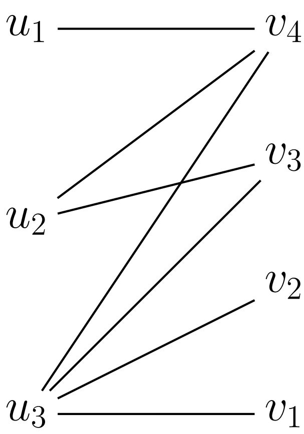


Figure 3.1: An example of a chain graph

to set inclusion. It is easily seen that this nesting property holds for U if and only if it holds for V. Figure 3.1 shows an example of a chain graph.

Now, as our terminology might suggest, the neighbourhood $K(a)$ of some player $a \in A$ in a tournament K coincides with the neighbourhood of the corresponding vertex in G_K. If G_K is a chain graph we can therefore enumerate A as $\{a_1, \ldots, a_m\}$ such that $K(a_i) \subseteq K(a_{i+1})$ for each $1 \leq i < m$. This indicates that each a_{i+1} has performed *at least as well* as a_i in a strong sense: every opponent which a_i defeated was also defeated by a_{i+1}, and a_{i+1} may have additionally defeated opponents which a_i did not.[4] It seems only natural in this case that one should rank a_i (weakly) below a_{i+1}. Appealing to transitivity and the fact that each $a \in A$ appears as *some* a_i, we see that any tournament K where G_K is a chain graph comes pre-equipped with a natural total preorder on A, where a' ranks higher than than a if and only if $K(a) \subseteq K(a')$. The duality of the neighbourhood-nesting property for chain graphs implies that B can also be totally preordered, with b' ranked higher than b if and only if $K^{-1}(b) \supseteq K^{-1}(b')$.[5] Moreover, these total preorders relate to the tournament results in an important sense: if a defeats b and b' ranks worse than b, then a must defeat b' also. That is, the neighbourhood of each $a \in A$ is *downwards closed* w.r.t the ranking of B, and the neighbourhood of each $b \in B$ is *upwards closed* in A.

Tournaments corresponding to chain graphs will be said to satisfy the *chain property*, and will accordingly be called *chain tournaments*. We give a simpler (but equivalent) definition which does not refer to the underlying graph G_K. First, define relations $\leqslant_K^{\mathcal{A}}, \leqslant_K^{\mathcal{B}}$ on A and B respectively by $a \leqslant_K^{\mathcal{A}} a'$ iff $K(a) \subseteq K(a')$ and $b \leqslant_K^{\mathcal{B}} b'$ iff $K^{-1}(b) \supseteq K^{-1}(b')$, for any tournament K.

[4]Note that this is a more robust notion of performance than comparing the neighbourhoods of a_i and a_{i+1} by *cardinality*, which may fail to account for differences in the strength of opponents when counting wins and losses.

[5]Note that the ordering of the Bs is reversed compared to the As, since the larger $K^{-1}(b)$ the *worse* b has performed.

Definition 3.1.4. *A tournament K has the* chain property *if $\leqslant_K^{\mathcal{A}}$ is a total preorder.*

According to the duality principle mentioned already, the chain property implies that $\leqslant_K^{\mathcal{B}}$ is also a total preorder. Note that the relations $\leqslant_K^{\mathcal{A}}$ and $\leqslant_K^{\mathcal{B}}$ are analogues of the *covering relation* for non-bipartite tournaments (Brandt, Brill, et al. 2016).

Example 3.1.1. *Consider $K = \begin{bmatrix} 1&0&0&0 \\ 1&1&0&0 \\ 1&1&1&1 \end{bmatrix}$. Then $K(1) \subset K(2) \subset K(3)$, so K has the chain property. In fact, K is the tournament corresponding to the chain graph G from Figure 3.1.*

3.2 Ranking via Chain Editing

We have seen that chain tournaments come equipped with natural rankings of A and B. Such tournaments represent an "ideal" situation, wherein the abilities of the players on both sides of the tournament are perfectly nested. Of course this may not be so in reality: the nesting may be broken by some $a \in A$ winning a match it ought not to by chance, or by losing a match by accident.

One idea for recovering a ranking in this case, originally suggested by Jiao, Ravi, and Gatterbauer (2017), is to apply *chain editing*: find the minimum number of edge changes required to convert the graph G_K into a chain graph. This process can be seen as correcting the "noise" in an observed tournament K to obtain an ideal ranking. In this section we introduce the class of operators producing rankings in this way.

3.2.1 Chain-minimal Operators

To define chain-editing in our framework we once again present an equivalent definition which does not refer to the underlying graph G_K: the number of edge changes between graphs can be replaced by the *Hamming distance* between tournament matrices.

Definition 3.2.1. *For $m, n \in \mathbb{N}$, let $\mathcal{C}_{m,n}$ denote the set of all $m \times n$ chain tournaments. For an $m \times n$ tournament K, write $\mathcal{M}(K) = \arg\min_{K' \in \mathcal{C}_{m,n}} d(K, K') \subseteq \mathcal{K}$ for the set of chain tournaments closest to K w.r.t the Hamming distance $d(K, K') = |\{(a, b) \in A \times B \mid K_{ab} \neq K'_{ab}\}|$. Let $m(K)$ denote this minimum distance.*

Note that chain editing, which is NP-hard in general (Jiao, Ravi, and Gatterbauer 2017), amounts to finding a single element of $\mathcal{M}(K)$.[6] We comment further on the computational complexity of chain editing in Section 3.6. The following property characterises chain editing-based operators T.

> **Chain-min**. For every tournament K there is $K' \in \mathcal{M}(K)$ such that $T(K) = (\leqslant^{\mathcal{A}}_{K'}, \leqslant^{\mathcal{B}}_{K'})$.

That is, the ranking of K is obtained by choosing the neighbourhood-subset rankings for some closest chain tournament K'. Operators satisfying **Chain-min** will be called *chain-minimal*.

Example 3.2.1. *Consider* $K = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}$. *$K$ does not have the chain property, since neither* $K(1) \subseteq K(2)$ *nor* $K(2) \subseteq K(1)$. *The set* $\mathcal{M}(K)$ *consists of four tournaments a distance of 2 from* K*:*

$$\mathcal{M}(K) = \left\{ \begin{bmatrix} 1 & \mathbf{1} & 1 & 0 \\ 1 & 1 & 0 & 0 \\ \mathbf{1} & 1 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & \mathbf{0} & 0 \\ 1 & 1 & 0 & 0 \\ \mathbf{1} & 1 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & \mathbf{0} & 0 & 0 \\ \mathbf{1} & 1 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & \mathbf{1} & 0 \\ \mathbf{1} & 1 & 1 & 1 \end{bmatrix} \right\}$$

The corresponding rankings are $(213, \{12\}34)$, $(123, 12\{34\})$, $(213, 13\{24\})$ *and* $(123, \{13\}24)$.[7]

Example 3.2.1 shows that there is no unique chain-minimal operator, since for a given tournament K there may be several closest chain tournaments to choose from. In Section 3.4 we introduce a principled way to single out a *unique* chain tournament and thereby construct a well-defined chain-minimal operator.

3.2.2 A Maximum Likelihood Interpretation

So far we have motivated **Chain-min** as a way to fix errors in a tournament and recover the ideal or *true* ranking. In this section we make this notion precise by defining a probabilistic model in which chain-minimal rankings arise as maximum likelihood estimates. The maximum likelihood approach has been applied for (non-bipartite) tournaments (e.g. the Bradley-Terry model (Bradley and Terry 1952; González-Díaz, Hendrickx, and Lohmann 2014)), voting in social choice theory (Elkind and Slinko 2016), truth discovery (D. Wang et al. 2012), belief merging (Everaere, Konieczny, and Marquis 2020) and other related problems.

In this approach we take an epistemic view of tournament ranking: it is assumed there exists a true "state of the world" which determines the tournament results along with objective rankings of A and B. A given tournament K is then seen as a *noisy observation* derived from the true state, and a *maximum likelihood estimate* is a state for which the probability of observing K is maximal.

[6] The decision problem associated with chain editing – which in tournament terms is the question of whether $m(K) \leq k$ for a given integer k – is NP-complete (Drange et al. 2015).

[7] Here $a_1 a_2 a_3$ is shorthand for the ranking $a_1 \prec a_2 \prec a_3$ of A, and similar for B. Elements in brackets are ranked equally.

More specifically, a state of the world is represented as a vector of *skill levels* for the players in A and B.[8]

Definition 3.2.2. *For a fixed size $m \times n$, a* state of the world *is a tuple $\theta = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$, where $\boldsymbol{x} \in \mathbb{R}^m$ and $\boldsymbol{y} \in \mathbb{R}^n$ satisfies the following properties:*

$$\forall a, a' \in A \quad (x_a < x_{a'} \implies \exists b \in B : x_a < y_b \leq x_{a'}) \tag{3.1}$$

$$\forall b, b' \in B \quad (y_b < y_{b'} \implies \exists a \in A : y_b \leq x_a < y_{b'}) \tag{3.2}$$

where $A = [m]$, $B = [n]$. Write $\Theta_{m,n}$ for the set of all $m \times n$ states.

For $a \in A$, x_a is the *skill level* of a in state θ (and similarly for y_b). These skill levels represent the true capabilities of the players in A and B in state θ: a is capable of defeating b if and only if $x_a \geq y_b$. Note that (3.1) suggests a simple form of *explainability*: a' can only be strictly more skilful than a if there is some $b \in B$ which *explains* this fact, i.e. some b which a' can defeat but a cannot ((3.2) is analogous for the Bs). These conditions are intuitive if we assume that skill levels are relative to the sets A and B currently under consideration (i.e. they do not reflect the abilities of players in future matches against new contenders outside of A or B). Finally note that our states of the world are *richer* than the output of an operator, in contrast to other work in the literature (Bradley and Terry 1952; González-Díaz, Hendrickx, and Lohmann 2014; Elkind and Slinko 2016). Specifically, a state θ contains extra information in the form of comparisons between A and B.

Noise is introduced in the observed tournament K via *false positives* (where $a \in A$ defeats a more skilled $b \in B$ by accident) and *false negatives* (where $a \in A$ is defeated by an inferior $b \in B$ by mistake).[9] The noise model is therefore parametrised by the false positive and false negative rates $\boldsymbol{\alpha} = \langle \alpha_+, \alpha_- \rangle \in [0, 1]^2$, which we assume are the same for all $a \in A$.[10] We also assume that noise occurs independently across all matches.

Definition 3.2.3. *Let $\boldsymbol{\alpha} = \langle \alpha_+, \alpha_- \rangle \in [0, 1]^2$. For each $m, n \in \mathbb{N}$ and $\theta = \langle \boldsymbol{x}, \boldsymbol{y} \rangle \in \Theta_{m,n}$, consider independent binary random variables X_{ab} representing the outcome of a match between $a \in [m]$ and $b \in [n]$, where*

$$P_{\boldsymbol{\alpha}}(X_{ab} = 1 \mid \theta) = \begin{cases} \alpha_+, & x_a < y_b \\ 1 - \alpha_-, & x_a \geq y_b \end{cases} \tag{3.3}$$

$$P_{\boldsymbol{\alpha}}(X_{ab} = 0 \mid \theta) = \begin{cases} 1 - \alpha_+, & x_a < y_b \\ \alpha_-, & x_a \geq y_b \end{cases} \tag{3.4}$$

[8] For simplicity we use numerical skill levels here, although it would suffice to have a partial preorder on $A \sqcup B$ such that each $a \in A$ is comparable with every $b \in B$.

[9] Note that a false positive for a is a false negative for b and vice versa.

[10] This is a strong assumption, and it may be more realistic to model the false positive/negative rates as a function of x_a. We leave this to future work.

This defines a probability distribution $P_{\boldsymbol{\alpha}}(\cdot \mid \theta)$ over $m \times n$ tournaments by

$$P_{\boldsymbol{\alpha}}(K \mid \theta) = \prod_{(a,b) \in [m] \times [n]} P_{\boldsymbol{\alpha}}(X_{ab} = K_{ab} \mid \theta)$$

Here $P_{\boldsymbol{\alpha}}(K \mid \theta)$ is the probability of observing the tournament results K when the false positive and negative rates are given by $\boldsymbol{\alpha}$ and the true state of the world is θ. Note that the four cases in (3.3) and (3.4) correspond to a false positive, true positive, true negative and false negative respectively. We can now define a maximum likelihood operator.

Definition 3.2.4. *Let $\boldsymbol{\alpha} \in [0,1]^2$ and $m, n \in \mathbb{N}$. Then $\theta \in \Theta_{m,n}$ is a* maximum likelihood estimate *(MLE) for an $m \times n$ tournament K w.r.t $\boldsymbol{\alpha}$ if $\theta \in \arg\max_{\theta' \in \Theta_{m,n}} P_{\boldsymbol{\alpha}}(K \mid \theta')$. An operator T is a* maximum likelihood operator *w.r.t $\boldsymbol{\alpha}$ if for any $m, n \in \mathbb{N}$ and any $m \times n$ tournament K there is an MLE $\theta = \langle \boldsymbol{x}, \boldsymbol{y} \rangle \in \Theta_{m,n}$ for K such that $a \preceq_K^T a'$ iff $x_a \leq x_{a'}$ and $b \sqsubseteq_K^T b'$ iff $y_b \leq y_{b'}$.*

To help analyse MLE operators, we consider the tournament K_θ associated with each state $\theta = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$, given by $[K_\theta]_{ab} = 1$ if $x_a \geq y_b$ and $[K_\theta]_{ab} = 0$ otherwise. Note that K_θ is the unique tournament with non-zero probability when there are no false positive or false negatives. The following technical lemma obtains an expression for $P_{\boldsymbol{\alpha}}(K \mid \theta)$ in terms of K_θ and K.

Lemma 3.2.1. *Let K be an $m \times n$ tournament, $\boldsymbol{\alpha} \in [0,1]^2$ and $\theta \in \Theta_{m,n}$. Then*

$$\begin{aligned} P_{\boldsymbol{\alpha}}(K \mid \theta) = \prod_{a \in A} \alpha_+^{|K(a) \setminus K_\theta(a)|} (1 - \alpha_-)^{|K(a) \cap K_\theta(a)|} \\ (1 - \alpha_+)^{|B \setminus (K(a) \cup K_\theta(a))|} \alpha_-^{|K_\theta(a) \setminus K(a)|}. \end{aligned}$$

Proof. Write $p_{ab,K}$ for $P_{\boldsymbol{\alpha}}(X_{ab} = K_{ab} \mid \theta)$. Expanding the product in Definition 3.2.3, we have

$$P_{\boldsymbol{\alpha}}(K \mid \theta) = \prod_{a \in A} \prod_{b \in B} p_{ab,K}.$$

Let $a \in A$. Note that B can be written as the disjoint union $B = B_1 \cup B_2 \cup B_3 \cup B_4$, where

$$\begin{aligned} B_1 &= K(a) \setminus K_\theta(a) \\ B_2 &= K(a) \cap K_\theta(a) \\ B_3 &= B \setminus (K(a) \cup K_\theta(a)) \\ B_4 &= K_\theta(a) \setminus K(a). \end{aligned}$$

Recall that $b \in K_\theta(a)$ iff $x_a \geq y_b$ (where $\theta = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$). It follows that

- $b \in B_1$ iff $K_{ab} = 1$ and $x_a < y_b$
- $b \in B_2$ iff $K_{ab} = 1$ and $x_a \geq y_b$

- $b \in B_3$ iff $K_{ab} = 0$ and $x_a < y_b$
- $b \in B_4$ iff $K_{ab} = 0$ and $x_a \geq y_b$

Note that this correspond exactly to the four cases in (3.3) and (3.4) which define $p_{ab,K}$; we have

$$p_{ab,K} = \begin{cases} \alpha_+, & b \in B_1 \\ 1 - \alpha_-, & b \in B_2 \\ 1 - \alpha_+, & b \in B_3 \\ \alpha_-, & b \in B_4. \end{cases}$$

Consequently

$$\begin{aligned} \prod_{b \in B} p_{ab,K} &= \left(\prod_{b \in B_1} \alpha_+ \right) \left(\prod_{b \in B_2} (1 - \alpha_-) \right) \left(\prod_{b \in B_3} (1 - \alpha_+) \right) \left(\prod_{b \in B_4} \alpha_- \right) \\ &= \alpha_+^{|B_1|} (1 - \alpha_-)^{|B_2|} (1 - \alpha_+)^{|B_3|} \alpha_-^{|B_4|} \\ &= \alpha_+^{|K(a) \setminus K_\theta(a)|} (1 - \alpha_-)^{|K(a) \cap K_\theta(a)|} \\ &\quad (1 - \alpha_+)^{|B \setminus (K(a) \cup K_\theta(a))|} \alpha_-^{|K_\theta(a) \setminus K(a)|}. \end{aligned}$$

Taking the product over all $a \in A$ we reach the desired expression for $P_{\boldsymbol{\alpha}}(K \mid \theta)$. ☐

Expressed in terms of K_θ, the MLEs take a particularly simple form if $\alpha_+ = \alpha_-$, i.e. if false positives and false negatives occur at the same rate.

Lemma 3.2.2. *Let $\boldsymbol{\alpha} = \langle \beta, \beta \rangle$ for some $\beta < \frac{1}{2}$. Then θ is an MLE for K if and only if $\theta \in \arg\min_{\theta' \in \Theta_{m,n}} d(K, K_{\theta'})$.*

Proof. Let K be an $m \times n$ tournament. By Lemma 3.2.1,

$$\begin{aligned} P_{\boldsymbol{\alpha}}(K \mid \theta) = \Bigg(\prod_{a \in A} & \alpha_+^{|K(a) \setminus K_\theta(a)|} (1 - \alpha_-)^{|K(a) \cap K_\theta(a)|} \\ & (1 - \alpha_+)^{|B \setminus (K(a) \cup K_\theta(a))|} \alpha_-^{|K_\theta(a) \setminus K(a)|} \Bigg). \end{aligned}$$

Plugging in $\alpha_+ = \alpha_- = \beta$ and simplifying, one can obtain

$$P_{\boldsymbol{\alpha}}(K \mid \theta) = c \prod_{a \in A} \left(\frac{\beta}{1 - \beta} \right)^{|K(a) \triangle K_\theta(a)|},$$

where $X \triangle Y = (X \setminus Y) \cup (Y \setminus X)$ is the symmetric difference of two sets X and Y, and $c = (1 - \beta)^{|A| \cdot |B|}$ is a positive constant that does not depend on θ. Now, $P_{\boldsymbol{\alpha}}(K \mid \theta)$ is positive, and is maximal when its logarithm is. We have

$$\begin{aligned} \log P_{\boldsymbol{\alpha}}(K \mid \theta) &= \log c + \log \left(\frac{\beta}{1 - \beta} \right) \sum_{a \in A} |K(a) \triangle K_\theta(a)| \\ &= \log c + \log \left(\frac{\beta}{1 - \beta} \right) d(K, K_\theta) \end{aligned}.$$

Since $\log c$ is constant and $\beta < 1/2$ implies $\log\left(\frac{\beta}{1-\beta}\right) < 0$, it follows that $\log P_{\boldsymbol{\alpha}}(K \mid \theta)$ is maximised exactly when $d(K, K_\theta)$ is minimised, which proves the result. □

This result characterises the MLE states for K as those for which K_θ is the closest to K. As it turns out, the tournaments K_θ that arise in this way are exactly those with the chain property.

Lemma 3.2.3. *Let $\theta = \langle \boldsymbol{x}, \boldsymbol{y} \rangle \in \Theta_{m,n}$. Then for all $a, a' \in A$ and $b, b' \in B$:*

1. *$K_\theta(a) \subseteq K_\theta(a')$ iff $x_a \leq x_{a'}$*
2. *$K_\theta^{-1}(b) \supseteq K_\theta^{-1}(b')$ iff $y_b \leq y_{b'}$.*

Proof. We prove (1); (2) is shown similarly. Let $a, a' \in A$. First suppose $x_a \leq x_{a'}$. Let $b \in K_\theta(a)$. Then $y_b \leq x_a \leq x_{a'}$, so $b \in K_\theta(a')$ also. This shows $K_\theta(a) \subseteq K_\theta(a')$.

Now suppose $K_\theta(a) \subseteq K_\theta(a')$. For the sake of contradiction, suppose $x_a > x_{a'}$. By (3.1) in the definition of a state (Definition 3.2.2), there is $b \in B$ such that $x_{a'} < y_b \leq x_a$. But this means $b \in K_\theta(a) \setminus K_\theta(a')$, which contradicts $K_\theta(a) \subseteq K_\theta(a')$. Thus (1) is proved. □

Lemma 3.2.4. *An $m \times n$ tournament K has the chain property if and only if $K = K_\theta$ for some $\theta \in \Theta_{m,n}$.*

Proof. The "if" direction follows from Lemma 3.2.3 (1): if $\theta = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$ and $a, a' \in A$ then either $x_a \leq x_{a'}$ – in which case $K_\theta(a) \subseteq K_\theta(a')$ – or $x_{a'} < x_a$ – in which case $K_\theta(a') \subseteq K_\theta(a)$. Therefore K_θ has the chain property.

For the "only if" direction, suppose K has the chain property. Define $\theta = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$ by

$$x_a = |\{a' \in A \mid K(a') \subseteq K(a)\}|$$
$$y_b = \begin{cases} \min\{x_a \mid a \in K^{-1}(b)\}, & K^{-1}(b) \neq \emptyset \\ 1 + |A|, & K^{-1}(b) = \emptyset \end{cases}$$

It is easily that since the neighbourhood-subset relation $\leqslant_K^A$ is a total preorder, we have $K(a) \subseteq K(a')$ if and only if $x_a \leq x_{a'}$. First we show that $K_\theta = K$ by showing that $K_{ab} = 1$ if and only if $[K_\theta]_{ab} = 1$. Suppose $K_{ab} = 1$. Then $a \in K^{-1}(b)$, so $y_b = \min\{x_{a'} \mid a' \in K^{-1}(b)\} \leq x_a$ and consequently $[K_\theta]_{ab} = 1$.

Now suppose $[K_\theta]_{ab} = 1$. Then $x_a \geq y_b$. We must have $K^{-1}(b) \neq \emptyset$; otherwise $y_b = 1 + |A| > |A| \geq x_a$. We can therefore take $\hat{a} \in \arg\min_{a' \in K^{-1}(b)} x_{a'}$. By definition of y_b, $x_{\hat{a}} = y_b \leq x_a$. But $x_{\hat{a}} \leq x_a$ implies $K(\hat{a}) \subseteq K(a)$; since $\hat{a} \in K^{-1}(b)$ this gives $b \in K(\hat{a})$ and $b \in K(a)$, i.e. $K_{ab} = 1$. This completes the claim that $K = K_\theta$.

It only remains to show that θ satisfies conditions (3.1) and (3.2) of Definition 3.2.2. For (3.1), suppose $x_a < x_{a'}$. Then $K(a) \subset K(a')$, i.e there is $b \in K(a') \setminus K(a) = K_\theta(a') \setminus K_\theta(a)$. But $b \in K_\theta(a')$ gives $y_b \leq x_{a'}$, and $b \notin K_\theta(a)$ gives $x_a < y_b$; this shows that (3.1) holds.

For (3.2), suppose $y_b < y_{b'}$. Clearly $K^{-1}(b) \neq \emptyset$ (otherwise $y_b = 1 + |A|$ is maximal). Thus there is $a \in K^{-1}(b)$ such that $y_b = x_a$. This of course means $x_a < y_{b'}$; in particular we have $y_b \leq x_a < y_{b'}$ as required for (3.2).

We have shown that $K = K_\theta$ and that $\theta \in \Theta_{m,n}$, and the proof is complete. □

Note that the proof of Lemma 3.2.3 relies crucially on (3.1) and (3.2) in the definition of a state. Combining all the results so far we obtain our first main result: the maximum likelihood operators for $\boldsymbol{\alpha} = \langle \beta, \beta \rangle$ are exactly the chain-minimal operators.

Theorem 3.2.1. *Let $\boldsymbol{\alpha} = \langle \beta, \beta \rangle$ for some $\beta < \frac{1}{2}$. Then T is a maximum likelihood operator w.r.t $\boldsymbol{\alpha}$ if and only if T satisfies* ***Chain-min****.*

Proof. First we show that for any $m, n \in \mathbb{N}$ and any $m \times n$ tournament K it holds that θ is an MLE state for K if and only if $K_\theta \in \mathcal{M}(K)$.

Indeed, fix some m, n and K. Write $\mathcal{K}_{\Theta_{m,n}} = \{K_\theta \mid \theta \in \Theta_{m,n}\}$. By Lemma 3.2.2, θ is an MLE if and only if $d(K, K_\theta) \leq d(K, K_{\theta'})$ for all $\theta' \in \Theta_{m,n}$, i.e. $K_\theta \in \arg\min_{K' \in \mathcal{K}_{\Theta_{m,n}}} d(K, K')$. But by Lemma 3.2.4, $\mathcal{K}_{\Theta_{m,n}}$ is just $\mathcal{C}_{m,n}$, the set of all $m \times n$ tournaments with the chain property. We see that $\arg\min_{K' \in \mathcal{K}_{\Theta_{m,n}}} d(K, K') = \arg\min_{K' \in \mathcal{C}_{m,n}} d(K, K') = \mathcal{M}(K)$ by definition of $\mathcal{M}(K)$. This shows that θ is an MLE iff $K_\theta \in \mathcal{M}(K)$.

Now, by definition, T satisfies **Chain-min** iff for every tournament K there is $K' \in \mathcal{M}(K)$ such that $T(K) = (\leqslant^{\mathcal{A}}_{K'}, \leqslant^{\mathcal{B}}_{K'})$. Using Lemma 3.2.4 and the above result, $K' \in \mathcal{M}(K)$ if and only if $K' = K_\theta$ for some MLE θ for K. We see that **Chain-min** can be equivalently stated as follows: for all tournament K there exists an MLE θ such that $T(K) = (\leqslant^{\mathcal{A}}_{K_\theta}, \leqslant^{\mathcal{B}}_{K_\theta})$. But by Lemma 3.2.3 we have $a \leqslant^{\mathcal{A}}_{K_\theta} a'$ iff $x_a \leq x_{a'}$ and $b \leqslant^{\mathcal{B}}_{K_\theta} b'$ iff $y_b \leq y_{b'}$ (where $\theta = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$). The above reformulation of **Chain-min** now coincides with the definition of a maximum likelihood operator, and we are done. □

Similar results can be obtained for other limiting values of $\boldsymbol{\alpha}$. If $\alpha_+ = 0$ and $\alpha_- \in (0, 1)$ then the MLE operators correspond to *chain completion*: finding the minimum number of edge *additions* required to make G_K a chain graph. This models situations where false positives never occur, although false negatives may (e.g. numerical entry questions in the case where A represents students and B exam questions (Jiao, Ravi, and Gatterbauer 2017)). Similarly, the case $\alpha_- = 0$ and $\alpha_+ \in (0, 1)$ corresponds to *chain deletion*, where edge additions are not allowed.

3.3 Axiomatic analysis

Chain-minimal operators have theoretical backing in a probabilistic sense due to the results of Section 3.2.2, but are they appropriate ranking methods in practise? To address this question we consider the *normative* properties of chain-minimal operators via the axiomatic method of social choice theory. We formulate several axioms for bipartite tournament ranking and assess whether they are compatible with **Chain-min**. It will be seen that an important *anonymity* axiom fails for all chain-minimal operators; later in Section 3.4 we describe a scenario in which this is acceptable and define a class of concrete operators for this case, and in Section 3.5 we relax the **Chain-min** requirement in order to gain anonymity.

3.3.1 The Axioms

We will consider five axioms – mainly adaptations of standard social choice properties to the bipartite tournament setting.

Symmetry Properties. We consider two symmetry properties. The first is a classic anonymity axiom, which says that an operator T should not be sensitive to the "labels" used to identify participants in a tournament. Axioms of this form are standard in social choice theory; a tournament version goes at least as far back as (Rubinstein 1980).

We need some notation: for a tournament K and permutations $\sigma : A \to A$, $\pi : B \to B$, let $\sigma(K)$ and $\pi(K)$ denote the tournament obtained by permuting the rows and columns of K by σ and π respectively, i.e. $[\sigma(K)]_{ab} = K_{\sigma^{-1}(a),b}$ and $[\pi(K)]_{ab} = K_{a,\pi^{-1}(b)}$. Note that in the statement of the axioms we omit universal quantification over K, $a, a' \in A$ and $b, b' \in B$ for brevity.

> **Anon**. Let $\sigma : A \to A$ and $\pi : B \to B$ be permutations. Then $a \preceq_K^T a'$ iff $\sigma(a) \preceq_{\pi(\sigma(K))}^T \sigma(a')$.

Our second axiom is specific to bipartite tournaments, and expresses a *duality* between the two sides A and B: given the two sets of conceptually disjoint entities participating in a bipartite tournament, it should not matter which one we label A and which one we label B. We need the notion of a *dual tournament.*

Definition 3.3.1. *The* dual tournament *of K is $\overline{K} = \mathbf{1} - K^\top$, where $\mathbf{1}$ denotes the matrix consisting entirely of 1s.*

$\overline{K}$ is essentially the same tournament as K, but with the roles of A and B swapped. In particular, $A_K = B_{\overline{K}}$, $B_K = A_{\overline{K}}$ and $K_{ab} = 1$ iff $\overline{K}_{ba} = 0$. Also note that $\overline{\overline{K}} = K$. The duality axiom states that the ranking of the Bs in K is the same as the As in $\overline{K}$.

> **Dual**. $b \sqsubseteq_K^T b'$ iff $b \preceq_{\overline{K}}^T b'$.

Whilst **Dual** is not necessarily a universally desirable property – one can imagine situations where A and B are not fully abstract and should not be treated symmetrically – it is important to consider in any study of bipartite tournaments. Note that **Dual** implies $a \preceq_K^T a'$ iff $a \sqsubseteq_K^T a'$, so that a **Dual**-operator can be defined by giving the ranking for one of A or B only, and defining the other by duality. This explains our choice to define **Anon** (and subsequent axioms) solely in terms of the A ranking: the analogous anonymity constraint for the B ranking follows from **Anon** together with **Dual**.

An Independence Property. Independence axioms play a crucial role in social choice. We present a bipartite adaptation of a classic axiom introduced in (Rubinstein 1980), which has subsequently been called *Independence of Irrelevant Matches* (González-Díaz, Hendrickx, and Lohmann 2014) in analogy with Independence of Irrelevant Alternatives in voting theory.

> **IIM**. If K_1, K_2 are tournaments of the same size with identical a-th and a'-th rows, then $a \preceq_{K_1}^T a'$ iff $a \preceq_{K_2}^T a'$.

IIM is a strong property, which says the relative ranking of a and a' does not depend on the results of any match not involving a or a'. This axiom has been questioned for generalised tournaments (González-Díaz, Hendrickx, and Lohmann 2014), and a similar argument can be made against it here: although each player in A faces the same opponents, we may wish to take the *strength* of opponents into account, e.g. by rewarding victories against highly-ranked players in B. Consequently we do not view **IIM** as an essential requirement, but rather introduce it to facilitate comparison with our work and the existing tournament literature.

Monotonicity Properties. Our final axioms are monotonicity properties, which express the idea that *more victories are better*. The first axiom follows our original intuition for constructing the natural ranking associated with a chain graph; namely that $K(a) \subseteq K(a')$ indicates a' has performed at least as well as a.

> **Mon**. If $K(a) \subseteq K(a')$ then $a \preceq_K^T a'$.

Note that **Mon** simply says $\preceq_K^T$ extends the (in general, partial) preorder $\leqslant_K^A$. Yet another standard axiom is positive responsiveness.

> **Pos-resp**. If $a \preceq_K^T a'$ and $K_{a',b} = 0$ for some $b \in B$, then $a \prec_{K+\mathbf{1}_{a',b}}^T a'$, where $\mathbf{1}_{a',b}$ is the matrix with 1 in position (a', b) and zeros elsewhere.

That is, adding an extra victory for a should only improve its ranking, with ties now broken in its favour. This version of positive responsiveness was again introduced in (Rubinstein 1980), where together with **Anon** and **IIM** it characterises the *points system* ranking method for round-robin tournaments, which simply ranks players according to the number of victories. The analogous operator in our framework is T_{count}, and it can be shown that T_{count} is uniquely characterised by **Anon**, **IIM**, **Pos-resp** and **Dual** (in fact, the proof follows the same argument as characterisation of voting in truth discovery in Theorem 2.3.3). Finally, note that **Pos-resp** also acts as a kind of *strategy-proofness*: a cannot improve its ranking by deliberately losing a match. Specifically, if $K_{ab} = 1$ and $a \preceq^T_K a'$, then **Pos-resp** implies $a \prec^T_{K-\mathbf{1}_{ab}} a'$.

3.3.2 Axiom Compatibility with Chain-min

We come to analysing the compatibility of **Chain-min** with the axioms. First, the negative results.

Theorem 3.3.1. *There is no operator satisfying* ***Chain-min*** *and any of* ***Anon****,* ***IIM*** *or* ***Pos-resp****.*

Proof. We take each axiom in turn. Let T be any operator satisfying **Chain-min**.

Anon. Consider $K = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$, and define permutations $\sigma = \pi = (1\,2)$, i.e. the permutations which simply swap 1 and 2. It is easily seen that $\pi(\sigma(K)) = K$. Supposing T satisfied **Anon**, we would get $1 \preceq^T_K 2$ iff $\sigma(1) \preceq^T_{\pi(\sigma(K))} \sigma(2)$ iff $2 \preceq^T_K 1$, which implies $1 \approx^T_K 2$. On the other hand, we have

$$\mathcal{M}(K) = \left\{\left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right], \left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right], \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right]\right\}$$

Since T satisfies **Chain-min** and $1, 2 \in A$ rank equally in $\preceq^T_K$, there must be $K' \in \mathcal{M}(K)$ such that 1 and 2 rank equally in $\leqslant^{\mathcal{A}}_{K'}$, i.e. $K'(1) = K'(2)$. But clearly there is no such K'; all tournaments in $\mathcal{M}(K)$ have distinct first and second rows. Hence T cannot satisfy **Anon**.

IIM. Suppose T satisfies **Chain-min** and **IIM**. Write

$$K_1 = \left[\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{smallmatrix}\right], \quad K_2 = \left[\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{smallmatrix}\right]$$

Note that the first and second rows of K_1 and K_2 are identical, so by **IIM** we have $1 \preceq^T_{K_1} 2$ iff $1 \preceq^T_{K_2} 2$. Both tournaments have a unique closest chain tournament requiring changes to only a single entry:

$$\mathcal{M}(K_1) = \left\{\left[\begin{smallmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{smallmatrix}\right]\right\}, \quad \mathcal{M}(K_2) = \left\{\left[\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{smallmatrix}\right]\right\}$$

Write $K_1{}'$ and $K_2{}'$ for these nearest chain tournaments respectively. By **Chain-min**, we must have $T(K_i) = (\leqslant^{\mathcal{A}}_{K_i{}'}, \leqslant^{\mathcal{B}}_{K_i{}'})$. In particular, $1 \prec^T_{K_1} 2$ and $2 \prec^T_{K_2} 1$. But this contradicts **IIM**, and we are done.

Pos-resp. Suppose T satisfies **Chain-min** and **Pos-resp**, and consider

$$K = \left[\begin{smallmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{smallmatrix}\right]$$

K has a unique closest chain tournament K':

$$\mathcal{M}(K) = \{K'\} = \left\{ \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \right\}$$

Chain-min therefore implies $T(K) = (\leqslant^{\mathcal{A}}_{K'}, \leqslant^{\mathcal{B}}_{K'})$. Note that $K'(1) = K'(2)$, so we have $1 \approx^T_K 2$. In particular, $1 \preceq^T_K 2$. Since $K_{23} = 0$, we may apply **Pos-resp** to get $1 \prec^T_{K+\mathbf{1}_{23}} 2$. But $K + \mathbf{1}_{23}$ is just K'. Since the chain property already holds for K', we have $\mathcal{M}(K') = \{K'\}$ and consequently

$$T(K + \mathbf{1}_{23}) = T(K') = (\leqslant^{\mathcal{A}}_{K'}, \leqslant^{\mathcal{B}}_{K'}) = T(K)$$

so in fact $1 \approx^T_{K+\mathbf{1}_{23}} 2$, contradicting **Pos-resp**. □

Note that the counterexample for **Anon** is particularly simple: we take $K = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$. Swapping the rows and columns brings us back to K, so **Anon** implies $1, 2 \in A$ rank equally. However, we saw that no chain tournament in $\mathcal{M}(K)$ yields this ranking.

The MLE results of Section 3.2.2 provides informal explanation for this result. For K above to arise in the noise model of Definition 3.2.3 there must have been two "mistakes" (false positives or false negatives). This is less likely than a single mistake from just one of $1, 2 \in A$, but the likelihood maximisation forces us to choose one or the other. A similar argument explains the **Pos-resp** failure.

It is also worth noting that **Anon** only fails at the last step of chain editing, where a single element of $\mathcal{M}(K)$ is chosen. Indeed, the set $\mathcal{M}(K)$ itself *does* exhibit the kind of symmetry one might expect: we have $\mathcal{M}(\pi(\sigma(K))) = \{\pi(\sigma(K')) \mid K' \in \mathcal{M}(K)\}$. This means that an operator which aggregates the rankings from *all* $K' \in \mathcal{M}(K)$ – e.g. any anonymous and neutral social welfare function (Zwicker 2016) – would satisfy **Anon**. The other axioms are compatible with **Chain-min**.

Theorem 3.3.2. *For each of **Dual** and **Mon**, there exists an operator satisfying **Chain-min** and the stated property.*

Despite the simplicity of **Mon**, Theorem 3.3.2 is deceptively difficult to prove, and we devote the rest of this section to its proof. We describe operators satisfying **Chain-def** and **Dual** or **Mon** non-constructively by first taking an *arbitrary* chain-minimal operator T, and using properties of the set $\mathcal{M}(K)$ to produce another operator T' satisfying **Dual** or **Mon**. Note also that we have not yet constructed an operator satisfying **Dual**, **Mon** and **Chain-min** simultaneously, although we conjecture that such operators do exist.

First we show compatibility of **Chain-min** and **Dual**. We need a preliminary result.

Lemma 3.3.1. *Let K be a tournament. Then*

1. $\leqslant^{\mathcal{B}}_K = \leqslant^{\mathcal{A}}_K$

2. $K' \in \mathcal{M}(K)$ *if and only if* $\overline{K'} \in \mathcal{M}\left(\overline{K}\right)$

Proof. Fix an $m \times n$ tournament K.

- Note that for any $b \in B$, we have $K^{-1}(b) = A \setminus \overline{K}(b)$. Indeed, for any $a \in A = A_K = B_{\overline{K}}$,

  $$\begin{aligned} a \in K^{-1}(b) &\iff K_{ab} = 1 \\ &\iff 1 - K_{ab} = 0 \\ &\iff \overline{K}_{ba} = 0 \\ &\iff a \notin \overline{K}(b) \end{aligned}$$

  This means that for any $b, b' \in B$,

  $$\begin{aligned} b \leqslant^{\mathcal{B}}_{K} b' &\iff K^{-1}(b) \supseteq K^{-1}(b') \\ &\iff A \setminus \overline{K}(b) \supseteq A \setminus \overline{K}(b') \\ &\iff \overline{K}(b) \subseteq \overline{K}(b') \\ &\iff b \leqslant^{\mathcal{A}}_{\overline{K}} b' \end{aligned}$$

  so $\leqslant^{\mathcal{B}}_{K} = \leqslant^{\mathcal{A}}_{\overline{K}}$.

- "only if": Suppose $K' \in \mathcal{M}(K)$. First we show that $\overline{K'}$ has the chain property. It is sufficient to show that $\leqslant^{\mathcal{B}}_{K'}$ is a total preorder,[11] since part (1) then implies $\leqslant^{\mathcal{A}}_{\overline{K'}}$ is a total preorder and $\overline{K'}$ has the chain property by definition.

  Since $\leqslant^{\mathcal{B}}_{K'}$ always has reflexivity and transitivity, we only need to show the totality property. Let $b, b' \in B$ and suppose $b \not\leqslant^{\mathcal{B}}_{K'} b'$. We must show $b' \leqslant^{\mathcal{B}}_{K'} b$, i.e. $(K')^{-1}(b') \supseteq (K')^{-1}(b)$. To that end, let $a \in (K')^{-1}(b)$.

  Since $(K')^{-1}(b) \not\supseteq (K')^{-1}(b')$, there is some $\hat{a} \in (K')^{-1}(b')$ with $\hat{a} \notin (K')^{-1}(b)$. That is, $b' \in K'(\hat{a})$ but $b \notin K'(\hat{a})$. Since $b \in K'(a)$, we have $K'(a) \nsubseteq K'(\hat{a})$. By the chain property for K', we get $K'(\hat{a}) \subset K'(a)$. Finally, this means $b' \in K'(\hat{a}) \subseteq K'(a)$, i.e $a \in (K')^{-1}(b')$. This shows $b' \leqslant^{\mathcal{B}}_{K'} b$ as required.

  It remains to show that $d(\overline{K}, \overline{K'})$ is minimal. Since every tournament is the dual of its dual, any $n \times m$ chain tournament is of the form $\overline{K''}$ for an $m \times n$ tournament K''. The above argument shows that the chain property is preserved by taking the dual, so that K'' has the chain property also. Since $K' \in \mathcal{M}(K)$, we have $d(K, K'') \geq d(K, K')$. It is easily verified that the Hamming distance is also preserved under duals, so

  $$d(\overline{K}, \overline{K'}) = d(K, K') \leq d(K, K'') = d(\overline{K}, \overline{K''})$$

[11] Note that we claim this holds for any K' with the chain property, but this has not yet been proven.

We have shown that $\overline{K'}$ is as close to $\overline{K}$ as any other $n \times m$ tournament with the chain property, which shows $\overline{K'} \in \mathcal{M}\left(\overline{K}\right)$ as required.

"if": Suppose $\overline{K'} \in \mathcal{M}\left(\overline{K}\right)$. By the "only if" statement above, we have $\overline{\overline{K'}} \in \mathcal{M}\left(\overline{\overline{K}}\right)$. But $\overline{\overline{K}} = K$ and $\overline{\overline{K'}} = K'$, so $K' \in \mathcal{M}(K)$ as required.

□

We can now find an operator with both **Chain-min** and **Dual**.

Proposition 3.3.1. *There exists an operator T satisfying **Chain-min** and **Dual**.*

Proof. Let T be an arbitrary operator satisfying **Chain-min**. Then there is a function $\alpha : \mathcal{K} \to \mathcal{K}$ such that $T(K) = (\leqslant^{\mathcal{A}}_{\alpha(K)}, \leqslant^{\mathcal{B}}_{\alpha(K)})$ and $\alpha(K) \in \mathcal{M}(K)$ for all tournaments K. We will construct a new function α', based on α, such that $\alpha'(\overline{K}) = \overline{\alpha'(K)}$.

Let $\ll$ be a total order on the set of all tournaments $\mathcal{K}$.[12] Write

$$T = \{K \in \mathcal{K} \mid K \ll \overline{K}\}$$

Note that since $K \neq \overline{K}$ for all K, exactly one of K and $\overline{K}$ lies in T. Informally, we view the tournaments in T as somehow "canonical", and those in $\mathcal{K} \setminus T$ as the dual of a canonical tournament. We use this notion to define α':

$$\alpha'(K) = \begin{cases} \alpha(K), & K \in T \\ \overline{\alpha(\overline{K})}, & K \notin T \end{cases}$$

First we claim $\alpha'(K) \in \mathcal{M}(K)$ for all K. Indeed, if $K \in T$ then $\alpha'(K) = \alpha(K) \in \mathcal{M}(K)$ by the assumption on α. Otherwise, $\alpha(\overline{K}) \in \mathcal{M}\left(\overline{K}\right)$, so Lemma 3.3.1 part (2) implies $\alpha'(K) = \overline{\alpha(\overline{K})} \in \mathcal{M}\left(\overline{\overline{K}}\right) = \mathcal{M}(K)$.

Next we show $\overline{\alpha'(K)} = \alpha'(\overline{K})$. First suppose $K \in T$. Then $\alpha'(K) = \alpha(K)$ and $\overline{K} \notin T$, so $\alpha'(\overline{K}) = \overline{\alpha(\overline{\overline{K}})} = \overline{\alpha(K)} = \overline{\alpha'(K)}$ as required. Similarly, if $K \notin T$ then $\overline{K} \in T$, so $\alpha'(\overline{K}) = \alpha(\overline{K})$, and $\alpha'(K) = \overline{\alpha(\overline{K})} = \overline{\alpha'(\overline{K})}$. Taking the dual of both sides, we get $\overline{\alpha'(K)} = \alpha'(\overline{K})$.

Finally, define a new operator T' by $T'(K) = (\leqslant^{\mathcal{A}}_{\alpha'(K)}, \leqslant^{\mathcal{B}}_{\alpha'(K)})$. Since $\alpha'(K) \in \mathcal{M}(K)$ for all K, T' satisfies **Chain-min**. Moreover, using Lemma 3.3.1 part (1) and the fact that $\overline{\alpha'(K)} = \alpha'(\overline{K})$, for any tournament K and $b, b' \in B$ we have

$$\begin{aligned} b \sqsubseteq^{T'}_{K} b' &\iff b \leqslant^{\mathcal{B}}_{\alpha'(K)} b' \\ &\iff b \leqslant^{\mathcal{A}}_{\overline{\alpha'(K)}} b' \\ &\iff b \leqslant^{\mathcal{A}}_{\alpha'(\overline{K})} b' \\ &\iff b \sqsubseteq^{T'}_{\overline{K}} b' \end{aligned}$$

which shows T' also satisfies **Dual**. $\square$

To find an operator satisfying **Chain-min** and **Mon**, we proceed in three stages. First, Lemma 3.3.2 shows that if $K(a_1) \subseteq K(a_2)$ and $K' \in \mathcal{M}(K)$ is some closest chain tournament with the reverse inclusion $K'(a_2) \subseteq K'(a_1)$, then swapping a_1 and a_2 in K' yields another closest chain tournament $K'' \in \mathcal{M}(K)$. Next, we show in Lemma 3.3.3 that by performing successive swaps in this way, we can find $K' \in \mathcal{M}(K)$ such that $K'(a_1) \subseteq K'(a_2)$ whenever $K(a_1) \subset K(a_2)$ (note the strict inclusion). Finally, we modify this K' in Lemma 3.3.4 to additionally satisfy $K'(a_1) = K'(a_2)$ whenever $K(a_1) = K(a_2)$. This shows that there always exist an element of $\mathcal{M}(K)$ extending the neighbourhood-subset relation $\leqslant_K^{\mathcal{A}}$, and consequently it is possible to satisfy **Chain-min** and **Mon** simultaneously.

Definition 3.3.2. *Let K be a tournament and $a_1, a_2 \in A$. We denote by $\mathsf{swap}(K; a_1, a_2)$ the tournament obtained by swapping the a_1 and a_2-th rows of K, i.e.*

$$[\mathsf{swap}(K; a_1, a_2)]_{ab} = \begin{cases} K_{a_1,b}, & a = a_2 \\ K_{a_2,b}, & a = a_1 \\ K_{a,b}, & a \notin \{a_1, a_2\} \end{cases}$$

Lemma 3.3.2. *Suppose $K(a_1) \subseteq K(a_2)$ and $K' \in \mathcal{M}(K)$ is such that $K'(a_2) \subseteq K'(a_1)$. Then $\mathsf{swap}(K'; a_1, a_2) \in \mathcal{M}(K)$.*

Proof. Write $K'' = \mathsf{swap}(K'; a_1, a_2)$. It is clear that K'' has the chain property since K' does. Since $K' \in \mathcal{M}(K)$, we have $d(K, K'') \geq d(K, K')$. We will show that $d(K, K'') \leq d(K, K')$ also, which implies $d(K, K'') = d(K, K') = m(K)$ and thus $K'' \in \mathcal{M}(K)$.

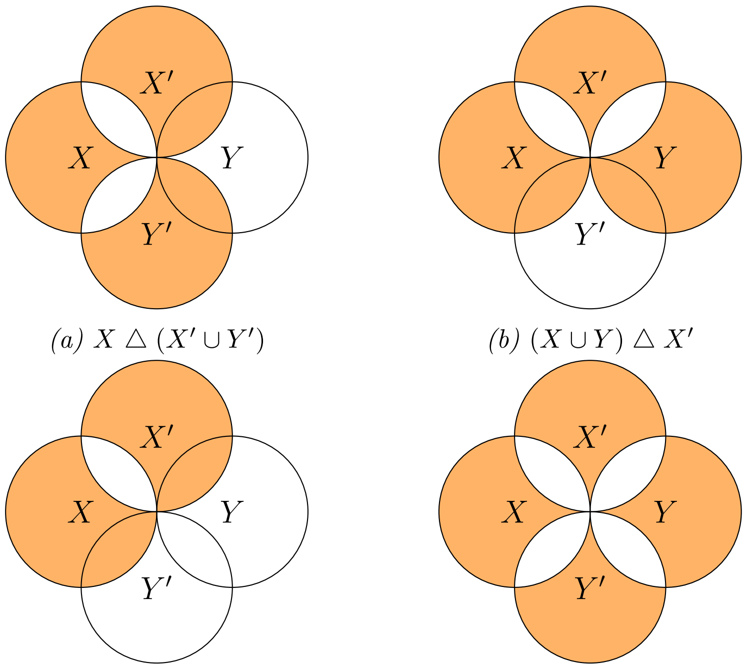
To that end, observe that for any tournament $\hat{K}$,

$$d(K, \hat{K}) = \sum_{a \in A} |K(a) \triangle \hat{K}(a)|$$

Noting that $K'(a) = K''(a)$ for $a \notin \{a_1, a_2\}$ and $K''(a_1) = K'(a_2)$, $K''(a_2) = K'(a_1)$, we have

$$\begin{aligned} d(K, K') - d(K, K'') &= \sum_{i \in \{1,2\}} (|K(a_i) \triangle K'(a_i)| - |K(a_i) \triangle K''(a_i)|) \\ &= |K(a_1) \triangle K'(a_1)| - |K(a_1) \triangle K'(a_2)| \\ &\quad + |K(a_2) \triangle K'(a_2)| - |K(a_2) \triangle K'(a_1)| \end{aligned}$$

[12]Note that $\mathcal{K}$ is countable, so such an order can be easily constructed. Alternatively, one could use the axiom of choice and appeal to the well-ordering theorem to obtain $\ll$.



(a) $X \triangle (X' \cup Y')$ *(b)* $(X \cup Y) \triangle X'$

(c) $X \triangle X'$ *(d)* $(X \cup Y) \triangle (X' \cup Y')$

Figure 3.2: Depictions of the sets in (3.5).

To simplify notation, write $X = K(a_1)$, $X' = K'(a_2)$, $Y = K(a_2) \setminus K(a_1)$ and $Y' = K'(a_1) \setminus K'(a_2)$. Since $K(a_1) \subseteq K(a_2)$ and $K'(a_2) \subseteq K'(a_1)$ by hypothesis, we have

$$\begin{aligned} K(a_1) &= X; \qquad & K(a_2) &= X \cup Y \\ K'(a_1) &= X' \cup Y'; \qquad & K'(a_2) &= X' \end{aligned}$$

and $X \cap Y = X' \cap Y' = \emptyset$. Rewriting the above we have

$$\begin{aligned} d(K, K') - d(K, K'') &= |K(a_1) \triangle K'(a_1)| + |K(a_2) \triangle K'(a_2)| \\ &\quad - |K(a_1) \triangle K'(a_2)| - |K(a_2) \triangle K'(a_1)| \\ &= |X \triangle (X' \cup Y')| + |(X \cup Y) \triangle X'| \\ &\quad - |X \triangle X'| - |(X \cup Y) \triangle (X' \cup Y')| \end{aligned} \tag{3.5}$$

Each of the symmetric differences in (3.5) are depicted in Figure 3.2. Note that each of these sets can be expressed as a union of the 8 disjoint subsets of $X \cup Y \cup X' \cup Y'$ shown in the figure. Expanding the symmetric differences in (3.5) and consulting Figure 3.2, it can be seen that most terms cancel out, and in fact we are left with

$$d(K, K') - d(K, K'') = 2|Y \cap Y'| \geq 0$$

This shows that $d(K, K'') \leq d(K, K')$, and the proof is complete. $\square$

We need some new notation. For a relation R on a set X and $x \in X$, write

$$U(x, R) = \{y \in X \mid x \; R \; y\}$$

$$L(x, R) = \{y \in X \mid y \; R \; x\}$$

for the upper- and lower-sets of x respectively.

Lemma 3.3.3. *For any tournament K there is $K' \in \mathcal{M}(K)$ such that for all $a \in A$:*

$$U(a, <_K^{\mathcal{A}}) \subseteq U(a, \leqslant_{K'}^{\mathcal{A}})$$

That is, $K(a) \subset K(a')$ implies $K'(a) \subseteq K'(a')$ for all $a, a' \in A$.

Proof. Write $A = \{a_1, \ldots, a_m\}$, ordered such that $|L(a_1, \leqslant_K^{\mathcal{A}})| \leq \cdots \leq |L(a_m, \leqslant_K^{\mathcal{A}})|$. We will show by induction that for each $0 \leq i \leq m$ there is $K_i \in \mathcal{M}(K)$ such that:

$$1 \leq j \leq i \implies U(a_j, <_K^{\mathcal{A}}) \subseteq U(a_j, \leqslant_{K_i}^{\mathcal{A}}) \tag{3.6}$$

The result follows by taking $K' = K_m$.

The case $i = 0$ is vacuously true, and we may take K_0 to be an arbitrary member of $\mathcal{M}(K)$. For the inductive step, suppose (3.6) holds for some $0 \leq i < m$. If $U(a_{i+1}, <_K^{\mathcal{A}}) = \emptyset$ then we may set $K_{i+1} = K_i$, so assume that $U(a_{i+1}, <_K^{\mathcal{A}})$ is non-empty. Take some $\hat{a} \in \min(U(a_{i+1}, <_K^{\mathcal{A}}), \leqslant_{K_i}^{\mathcal{A}})$. Then $\hat{a}$ has (one of) the smallest neighbourhoods in K_i amongst those in A with a strictly larger neighbourhood than a_{i+1} in K.

If $K_i(a_{i+1}) \subseteq K_i(\hat{a})$ then we claim (3.6) holds with $K_{i+1} = K_i$. Indeed, for $j < i + 1$ the inclusion in (3.6) holds since it does for K_i. For $j = i + 1$, let $a \in U(a_{i+1}, <_K^{\mathcal{A}})$. The definition of $\hat{a}$ implies $K_i(a) \not\subset K_i(\hat{a})$; since K_i has the chain property this means $K_i(\hat{a}) \subseteq K_i(a)$. Consequently $K_i(a_{i+1}) \subseteq K_i(\hat{a}) \subseteq K_i(a)$, i.e. $a \in U(a_{i+1}, \leqslant_{K_i}^{\mathcal{A}}) = U(a_{i+1}, \leqslant_{K_{i+1}}^{\mathcal{A}})$ as required.

For the remainder of the proof we therefore suppose $K_i(a_{i+1}) \not\subseteq K_i(\hat{a})$. The chain property for K_i gives $K_i(\hat{a}) \subset K_i(a_{i+1})$. Since $K_i \in \mathcal{M}(K)$ and $K(a_{i+1}) \subset K(\hat{a})$, we may apply Lemma 3.3.2. Set $K_{i+1} = \mathsf{swap}(K_i; a_{i+1}, \hat{a}) \in \mathcal{M}(K)$. The inclusion in (3.6) is easy to show for $j = i+1$: if $a \in U(a_{i+1}, <_K^{\mathcal{A}})$ then either $a = \hat{a}$ – in which case $K_{i+1}(a_{i+1}) \subset K_{i+1}(a)$ by construction – or $a \neq \hat{a}$ and $K_{i+1}(a_{i+1}) = K_i(\hat{a}) \subseteq K_i(a) = K_{i+1}(a)$. In either case $a \in U(a_{i+1}, \leqslant_{K_{i+1}}^{\mathcal{A}})$ as required.

Now suppose $1 \leq j < i + 1$. First note that due to our assumption on the ordering of $\{a_1, \ldots, a_m\}$, we have $a_j \neq \hat{a}$ (indeed, if $a_j = \hat{a}$ then $K(a_{i+1}) \subset K(a_j)$ and $|L(a_j, <_K^{\mathcal{A}})| > |L(a_{i+1}, <_K^{\mathcal{A}})|$). Since $a_j \neq a_{i+1}$ also, a_j was not involved in the swapping in the construction of K_{i+1}, and consequently $K_{i+1}(a_j) = K_i(a_j)$. Let $a \in U(a_j, <_K^{\mathcal{A}})$. We must show that $K_{i+1}(a_j) \subseteq K_{i+1}(a)$. We consider cases.

Case 1: $a = \hat{a}$. Using the fact that (3.6) holds for K_i we have

$$K_{i+1}(a_j) = K_i(a_j) \subseteq K_i(\hat{a}) \subset K_i(a_{i+1}) = K_{i+1}(\hat{a})$$

Case 2: $a = a_{i+1}$. Here $K(a_j) \subset K(a_{i+1}) \subset K(\hat{a})$, i.e. $\hat{a} \in U(a_j, <^{\mathcal{A}}_{K})$. Applying the inductive hypothesis again we have

$$K_{i+1}(a_j) = K_i(a_j) \subseteq K_i(\hat{a}) = K_{i+1}(a_{i+1})$$

Case 3: $a \notin \{\hat{a}, a_{i+1}\}$. Here neither a_j nor a were involved in the swap, so $K_{i+1}(a_j) = K_i(a_j) \subseteq K_i(a) = K_{i+1}(a)$.

By induction, the proof is complete. □

Lemma 3.3.4. *Let K be a tournament and suppose $K' \in \mathcal{M}(K)$ is such that $U(a, <^{\mathcal{A}}_{K}) \subseteq U(a, \leqslant^{\mathcal{A}}_{K'})$ for all $a \in A$. Then there is $K'' \in \mathcal{M}(K)$ such that $\leqslant^{\mathcal{A}}_{K} \subseteq \leqslant^{\mathcal{A}}_{K''}$.*

Proof. Let $A_1, \ldots, A_t \subseteq A$ be the equivalence classes of $\approx^{\mathcal{A}}_{K}$, the symmetric part of $\leqslant^{\mathcal{A}}_{K}$. Note that $a \approx^{\mathcal{A}}_{K} a'$ iff $K(a) = K(a')$, so we can associate each A_i with a neighbourhood $B_i \subseteq B$ such that $K(a) = B_i$ whenever $a \in A_i$.

Our aim is to select a single element from each equivalence class A_i, which we denote by $f(A_i)$, and modify K' to set the neighbourhood of each $a \in A_i$ to $K'(f(A_i))$. To that end, construct a function $f : \{A_1, \ldots, A_t\} \to A$ such that

$$f(A_i) \in \arg\min_{a \in A_i} |B_i \mathbin{\triangle} K'(a)| \in A_i$$

Define K'' by $K''_{ab} = K'_{f([a]),b}$, where $[a]$ denotes the equivalence class of a. Then $K''(a) = K'(f([a]))$ for all a.

Next we show that $K'' \in \mathcal{M}(K)$. Note that K'' has the chain property, since $a_1 \leqslant^{\mathcal{A}}_{K''} a_2$ iff $f([a_1]) \leqslant^{\mathcal{A}}_{K'} f([a_2])$, and $f([a_1]), f([a_2])$ are guaranteed to be comparable with respect to $\leqslant^{\mathcal{A}}_{K'}$ since K' has the chain property. To show $d(K, K'')$ is minimal, observe that

$$\begin{aligned} d(K, K'') &= \sum_{a \in A} |K(a) \mathbin{\triangle} K''(a)| \\ &= \sum_{i=1}^{t} \sum_{a \in A_i} |B_i \mathbin{\triangle} K'(f(A_i))| \end{aligned}$$

By definition of f, we have $|B_i \mathbin{\triangle} K'(f(A_i))| \leq |B_i \mathbin{\triangle} K'(a)|$ for all $a \in A_i$. Consequently

$$\begin{aligned} d(K, K'') &\leq \sum_{i=1}^{t} \sum_{a \in A_i} |B_i \mathbin{\triangle} K'(a)| \\ &= d(K, K') \\ &= m(K) \end{aligned}$$

which implies $K'' \in \mathcal{M}(K)$.

We are now ready to prove the result. Suppose $a \leqslant^{\mathcal{A}}_{K} a'$ i.e. $K(a) \subseteq K(a')$. If $K(a) = K(a')$ then $[a] = [a']$, so

$$K''(a) = K'(f([a])) = K'(f([a'])) = K''(a')$$

and in particular $K''(a) \subseteq K''(a')$. If instead $K(a) \subset K(a')$, then $K(f([a])) = K(a) \subset K(a') = K(f([a']))$, i.e. $f([a]) <^{\mathcal{A}}_{K} f([a'])$. By the assumption on K' in the statement of the lemma, this means $f([a]) \leqslant^{\mathcal{A}}_{K'} f([a'])$, and so

$$K''(a) = K'(f([a])) \subseteq K'(f([a'])) = K''(a')$$

In either case $K''(a) \subseteq K''(a')$, i.e. $a \leqslant^{\mathcal{A}}_{K''} a'$. Since a, a' were arbitrary, this shows that $\leqslant^{\mathcal{A}}_{K} \subseteq \leqslant^{\mathcal{A}}_{K''}$ as required. ☐

The pieces are now in place to prove the following.

Proposition 3.3.2. *There exists an operator T satisfying **Chain-min** and **Mon**.*

Proof. For any tournament K, write

$$\mathcal{M}_{\mathsf{mon}}(K) = \{K' \in \mathcal{M}(K) \mid \leqslant^{\mathcal{A}}_{K} \subseteq \leqslant^{\mathcal{A}}_{K'}\}$$

By Lemma 3.3.3 and Lemma 3.3.4, $\mathcal{M}_{\mathsf{mon}}(K)$ is non-empty. Let $\ll$ be any total order on the set $\mathcal{K}$ of all tournaments. Define a function $\alpha : \mathcal{K} \to \mathcal{K}$ by

$$\alpha(K) = \min(\mathcal{M}_{\mathsf{mon}}(K), \ll) \in \mathcal{M}_{\mathsf{mon}}(K)$$

Note that the minimum is unique since $\ll$ is a total order. Defining an operator T by $T(K) = (\leqslant^{\mathcal{A}}_{\alpha(K)}, \leqslant^{\mathcal{B}}_{\alpha(K)})$, we see that T satisfies **Chain-min** and **Mon**, as required. ☐

Theorem 3.3.2 now follows from Proposition 3.3.1 and Proposition 3.3.2.

3.4 Match-preference operators

The counterexample for **Chain-min** and **Anon** suggests that chain-minimal operators require some form of tie-breaking mechanism when the tournaments in $\mathcal{M}(K)$ cannot be distinguished while respecting anonymity. While this limits the use of chain-minimal operators as general purpose ranking methods, it is not such a problem if additional information is available to guide the tie-breaking. In this section we introduce a new class of operators for this case.

The core idea is to single out a unique chain tournament close to K by paying attention to not only the *number* of entries in K that need to be changed to produce a chain tournament, but *which* entries. Specifically, we assume the availability of a total order on the set of matrix indices $\mathbb{N} \times \mathbb{N}$ (the *matches*) which indicates our willingness to change an entry in K: the higher up (a, b) is in the ranking, the more acceptable it is to change K_{ab} during chain editing.

This total order – called the *match-preference relation* – is fixed for all tournaments K; this means we are dealing with extra information about how tournaments are *constructed in matrix form*, not extra information about any specific tournament K.

One possible motivation for such a ranking comes from cases where matches occur at distinct points in time. In this case the matches occurring more recently are (presumably) more representative of the players' *current* abilities, and we should therefore prefer to modify the outcome of old matches where possible.

For the formal definition we need notation for the *vectorisation* of a tournament K: for a total order $\trianglelefteq$ on $\mathbb{N}\times\mathbb{N}$ and an $m\times n$ tournament K, we write $\mathrm{vec}_{\trianglelefteq}(K)$ for the vector in $\{0,1\}^{mn}$ obtained by collecting the entries of K in the order given by $\trianglelefteq \restriction (A\times B)$,[13] starting with the minimal entry. That is, $\mathrm{vec}_{\trianglelefteq}(K) = (K_{a_1,b_1}, \dots, K_{a_{mn},b_{mn}})$, where $(a_1,b_1),\dots,(a_{mn},b_{mn})$ is the unique enumeration of $A\times B$ such that $(a_i,b_i)\trianglelefteq(a_{i+1},b_{i+1})$ for each i.

The operator corresponding to $\trianglelefteq$ is defined using the notion of a *choice function*: a function α which maps any tournament K to an element of $\mathcal{M}(K)$. Any such function defines a chain-minimal operator T by setting $T(K) = (\leqslant^{\mathcal{A}}_{\alpha(K)}, \leqslant^{\mathcal{B}}_{\alpha(K)})$.

Definition 3.4.1. *Let $\trianglelefteq$ be a total order on $\mathbb{N}\times\mathbb{N}$. Define an operator $T_{\trianglelefteq}$ according to the choice function*

$$\alpha_{\trianglelefteq}(K) = \underset{K'\in\mathcal{M}(K)}{\arg\min}\, \mathrm{vec}_{\trianglelefteq}(K\ominus K') \tag{3.7}$$

where $[K\ominus K']_{ab} = |K_{ab} - K'_{ab}|$, and the minimum is taken w.r.t the lexicographic ordering on $\{0,1\}^{|A|\cdot|B|}$.[14] *Operators generated in this way will be called* match-preference operators*.*

Example 3.4.1. *Let $\trianglelefteq$ be the lexicographic order*[15] *on $\mathbb{N}\times\mathbb{N}$ so that $\mathrm{vec}_{\trianglelefteq}(K\ominus K')$ is obtained by collecting the entries of $K\ominus K'$ row-by-row, from top to bottom and left to right. Take K from Example 3.2.1. Writing $K_1,\dots,K_4$ for the elements of $\mathcal{M}(K)$ in the order that they appear in Example 3.2.1 and setting $v_i = \mathrm{vec}_{\trianglelefteq}(K\ominus K_i)$, we have*

$$v_1 = (0\mathbf{1}00\ 0000\ \mathbf{1}0000); \qquad v_2 = (00\mathbf{1}0\ 0000\ \mathbf{1}0000)$$
$$v_3 = (0000\ 0\mathbf{1}00\ \mathbf{1}0000); \qquad v_4 = (0000\ 00\mathbf{1}0\ \mathbf{1}0000)$$

The lexicographic minimum is the one with the 1 entries as far right as possible, which in this case is v_4. Consequently $T_{\trianglelefteq}$ ranks K according to K_4, i.e. $1 \prec^{T_{\trianglelefteq}}_K 2 \prec^{T_{\trianglelefteq}}_K 3$ and $1 \approx^{T_{\trianglelefteq}}_K 3 \sqsubset^{T_{\trianglelefteq}}_K 2 \sqsubset^{T_{\trianglelefteq}}_K 4$.

To conclude the discussion of match-preference operators, we note that one can compute $\alpha_{\trianglelefteq}(K)$ as the unique closest chain tournament to K w.r.t a *weighted* Hamming distance, and thereby avoid the need to enumerate $\mathcal{M}(K)$ in full as per (3.7). First, a technical result is required.

[13] This denotes the restriction of $\trianglelefteq$ to $A\times B$, i.e. $\trianglelefteq\cap(A\times B)^2$.

[14] Note that $K\ominus K'$ is 1 in exactly the entries where K and K' differ.

[15] That is, $(a,b)\trianglelefteq(a',b')$ iff $a<a'$ or ($a=a'$ and $b\le b'$).

Lemma 3.4.1. *Let k and l be integers with $1 \leq k \leq l$. Then*

$$\sum_{i=k}^{l} 2^{-i} < 2^{-(k-1)}.$$

Proof. This follows from the formula for the sum of a finite geometric series:

$$\sum_{i=0}^{n-1} r^i = \frac{1-r^n}{1-r}$$

which holds for all $r \neq 1$. In this case we have

$$\begin{aligned}
\sum_{i=k}^{l} 2^{-i} &= \sum_{i=0}^{l} 2^{-i} - \sum_{i=0}^{k-1} 2^{-i} \\
&= \sum_{i=0}^{l} \left(\frac{1}{2}\right)^i - \sum_{i=0}^{k-1} \left(\frac{1}{2}\right)^i \\
&= \frac{1-\left(\frac{1}{2}\right)^{l+1}}{1-\left(\frac{1}{2}\right)} - \frac{1-\left(\frac{1}{2}\right)^{k}}{1-\left(\frac{1}{2}\right)} \\
&= 2\left(2^{-k} - 2^{-(l+1)}\right) \\
&= 2^{-(k-1)} - \underbrace{2^{-l}}_{>0} \\
&< 2^{-(k-1)}
\end{aligned}$$

as required. □

The characterisation in terms of weighted Hamming distances is as follows

Theorem 3.4.1. *Let $\trianglelefteq$ be a total order on $\mathbb{N} \times \mathbb{N}$. Then for any $m, n \in \mathbb{N}$ there exists a function $w : [m] \times [n] \to \mathbb{R}_{\geq 0}$ such that for all $m \times n$ tournaments K:*

$$\arg\min_{K' \in \mathcal{C}_{m,n}} d_w(K, K') = \{\alpha_{\trianglelefteq}(K)\} \tag{3.8}$$

where $d_w(K, K') = \sum_{(a,b) \in [m] \times [n]} w(a,b) \cdot |K_{ab} - K'_{ab}|$.

Proof. Let $\trianglelefteq$ be a total order on $\mathbb{N} \times \mathbb{N}$ and let $m, n \in \mathbb{N}$. For $a \in [m]$ and $b \in [n]$, write

$$p(a,b) = 1 + |\{(a', b') \in [m] \times [n] : (a', b') \triangleleft (a,b)\}|$$

for the "position" of (a,b) in $\trianglelefteq \restriction ([m] \times [n])$ (where 1 corresponds to the minimal pair). Define w by

$$w(a,b) = 1 + 2^{-p(a,b)}$$

If we abuse notation slightly and view w as an $m \times n$ matrix, we have, by construction, $\mathrm{vec}_{\trianglelefteq}(w) = (1+2^{-1}, \ldots, 1+2^{-mn})$. Noting that $|K_{ab} - K'_{ab}| = [K \ominus K']_{ab}$ for any tournaments K, K', and letting $\bullet$ denote the dot product, it is easy to see that

$$\begin{aligned} d_w(K, K') &= \mathrm{vec}_{\trianglelefteq}(w) \bullet \mathrm{vec}_{\trianglelefteq}(K \ominus K') \\ &= (1+2^{-1}, \ldots, 1+2^{-mn}) \bullet \mathrm{vec}_{\trianglelefteq}(K \ominus K') \\ &= d(K, K') + \boldsymbol{x} \bullet \mathrm{vec}_{\trianglelefteq}(K \ominus K') \end{aligned}$$

where $\boldsymbol{x} = (2^{-1}, \ldots, 2^{-mn})$ and $d(K, K')$ is the unweighted Hamming distance. In particular, since $\boldsymbol{x}$ and $\mathrm{vec}_{\trianglelefteq}(K \ominus K')$ are non-negative, we have $d_w(K, K') \geq d(K, K')$.

Now, we will show that for any $m \times n$ tournament K and $K' \in \mathcal{C}_{m,n}$ with $K' \neq \alpha_{\trianglelefteq}(K)$ we have $d_w(K, \alpha_{\trianglelefteq}(K)) < d_w(K, K')$. Since $\alpha_{\trianglelefteq}(K) \in \mathcal{M}(K) \subseteq \mathcal{C}_{m,n}$ by definition, this will show that $\alpha_{\trianglelefteq}(K)$ is the unique minimum in (3.8), as required.

So, let K be an $m \times n$ tournament and $K' \in \mathcal{C}_{m,n}$. To ease notation, write $v = \mathrm{vec}_{\trianglelefteq}(K \ominus \alpha_{\trianglelefteq}(K))$ and $v' = \mathrm{vec}_{\trianglelefteq}(K \ominus K')$. There are two cases.

Case 1: $K' \notin \mathcal{M}(K)$. In this case we have $d(K, K') \geq m(K) + 1$, and

$$\begin{aligned} d_w(K, \alpha_{\trianglelefteq}(K)) &= \underbrace{d(K, \alpha_{\trianglelefteq}(K))}_{=m(K)} + \boldsymbol{x} \bullet v \\ &= m(K) + \sum_{i=1}^{mn} 2^{-i} \cdot \underbrace{v_i}_{\leq 1} \\ &\leq m(K) + \underbrace{\sum_{i=1}^{mn} 2^{-i}}_{<2^{-0}=1} \\ &< m(K) + 1 \\ &\leq d(K, K') \\ &\leq d_w(K, K') \end{aligned}$$

where Lemma 3.4.1 was applied in the 4th step. This shows $d_w(K, \alpha_{\trianglelefteq}(K)) < d_w(K, K')$, as required.

Case 2: $K \in \mathcal{M}(K)$. In this case we have

$$\begin{aligned} d(K, \alpha_{\trianglelefteq}(K)) - d(K, K') &= (m(K) + \boldsymbol{x} \bullet v) - (m(K) + \boldsymbol{x} \bullet v') \\ &= \boldsymbol{x} \bullet (v - v') \end{aligned}$$

Now, since $K' \in \mathcal{M}(K)$, v' appears as one of the vectors over which the arg min is taken in (3.7). By definition of $\alpha_{\trianglelefteq}$ we therefore know that v strictly precedes v' with respect to the lexicographic order on $\{0, 1\}^{mn}$. Consequently there is $j \geq 1$ such that $v_i = v'_i$ for $i < j$ and $v_j < v'_j$. That is, $v_j = 0$ and

$v'_j = 1$. This means

$$
\begin{aligned}
d(K, \alpha_{\trianglelefteq}(K)) - d(K, K') &= \boldsymbol{x} \bullet (v - v') \\
&= \sum_{i=1}^{mn} 2^{-i}(v_i - v'_i) \\
&= \sum_{i=1}^{j-1} 2^{-i} \underbrace{(v_i - v'_i)}_{=0} + \sum_{i=j}^{mn} 2^{-i}(v_i - v'_i) \\
&= 2^{-j} \underbrace{(v_j - v'_j)}_{=-1} + \sum_{i=j+1}^{mn} 2^{-i} \underbrace{(v_i - v'_i)}_{\leq 1} \\
&\leq -2^{-j} + \sum_{i=j+1}^{mn} 2^{-i} \\
&< -2^{-j} + 2^{-j} \\
&= 0
\end{aligned}
$$

where Lemma 3.4.1 was applied in the second to last step. Again, this shows $d_w(K, \alpha_{\trianglelefteq}(K)) < d_w(K, K')$, and the proof is complete. □

For example, the weights corresponding to $\trianglelefteq$ from Example 3.4.1 and $m = 2$, $n = 3$ are $w = \left[\begin{smallmatrix} 1.5 & 1.25 & 1.125 \\ 1.0625 & 1.03125 & 1.015625 \end{smallmatrix}\right]$.

3.5 Relaxing chain-min

Having studied chain-minimal operators in some detail, we turn to two remaining problems: **Chain-min** is incompatible with **Anon**, and computing a chain-minimal operator is NP-hard. In this section we obtain both anonymity and tractability by relaxing the **Chain-min** requirement to a property we call *chain-definability*. We go on to characterise the class of operators with this weaker property via a greedy approximation algorithm, single out a particularly intuitive instance, revisit the axioms of Section 3.3, and present new axioms which characterise this intuitive instance.

3.5.1 Chain-definability

The source of the difficulties with **Chain-min** lies in the minimisation aspect of chain editing. A natural way to retain the spirit of **Chain-min** without the complications is to require that $T(K)$ corresponds to *some* chain tournament, not necessarily one closest to K. We call this property *chain-definability*.

> **Chain-def**. For every $m \times n$ tournament K there is $K' \in \mathcal{C}_{m,n}$ such that $T(K) = (\leqslant^{\mathcal{A}}_{K'}, \leqslant^{\mathcal{B}}_{K'})$.

Clearly **Chain-min** implies **Chain-def**. "Chain-definable" operators can also be cast in the MLE framework of Section 3.2.2 as those whose rankings correspond to *some* (not necessarily MLE) state θ.

At first glance it may seem difficult to determine whether a given pair of rankings correspond to a chain tournament, since the number of such tournaments grows rapidly with m and n. Fortunately, **Chain-def** can be characterised without reference to chain tournaments by considering the number of *ranks* of $\preceq^T_K$ and $\sqsubseteq^T_K$. In what follows $\mathsf{ranks}(\preceq)$ denotes the number of ranks of a total preorder $\preceq$, i.e. the number of equivalence classes of its symmetric part.

Theorem 3.5.1. *T satisfies **Chain-def** if and only if $|\mathsf{ranks}(\preceq^T_K) - \mathsf{ranks}(\sqsubseteq^T_K)| \leq 1$ for every tournament K.*

Proof. First we set up some notation. For a total preorder $\preceq$ on a set Z and $z \in Z$, write $[z]_\preceq$ for the rank of $\preceq$ containing z, i.e. the equivalence class of z in the symmetric closure of $\preceq$:

$$[z]_\preceq = \{z' \in Z \mid z \preceq z' \text{ and } z' \preceq z\}.$$

Also note that $\preceq$ can be extended to a total order on the ranks by setting $[z]_\preceq \leq [z']_\preceq$ iff $z \preceq z'$.

We start with the "only if" statement of the theorem. Suppose T satisfies **Chain-def**, and let K be a tournament. We need to show that $|\mathsf{ranks}(\preceq^T_K) - \mathsf{ranks}(\sqsubseteq^T_K)| \leq 1$.

By chain-definability, there is K' with the chain property such that $a \preceq^T_K a'$ iff $K'(a) \subseteq K'(a')$ and $b \sqsubseteq^T_K b'$ iff $(K')^{-1}(b) \supseteq (K')^{-1}(b')$. Write

$$\mathcal{X} = \{[a]_{\preceq^T_K} \mid a \in A, K'(a) \neq \emptyset\}$$

$$\mathcal{Y} = \{[b]_{\sqsubseteq^T_K} \mid b \in B, (K')^{-1}(b) \neq \emptyset\}$$

for the set of ranks in each of the two orders, excluding those who have empty neighbourhoods in K'. Note that $[a]_{\preceq^T_K} = [a']_{\preceq^T_K}$ if and only if $K'(a) = K'(a')$ (and similar for B).

We will show that $|\mathcal{X}| = |\mathcal{Y}|$. Enumerate $\mathcal{X} = \{X_1, \ldots, X_s\}$ and $\mathcal{Y} = \{Y_1, \ldots, Y_t\}$, ordered such that $X_1 < \cdots < X_s$ and $Y_1 < \cdots < Y_t$. First we show $|\mathcal{X}| \leq |\mathcal{Y}|$.

For each $1 \leq i \leq s$, the a_i be an arbitrary element of X_i. Then $a_1 \prec^T_K \cdots \prec^T_K a_s$, so $\emptyset \subset K'(a_1) \subset \cdots \subset K'(a_s)$. Since these inclusions are strict, we can choose $b_1, \ldots, b_s \in B$ such that $b_1 \in K'(a_1)$ and $b_{i+1} \in K'(a_{i+1}) \setminus K'(a_i)$ for $1 \leq i < s$.

It follows that $a_i \in (K')^{-1}(b_i) \setminus (K')^{-1}(b_{i+1})$, and thus $(K')^{-1}(b_i) \not\subseteq (K')^{-1}(b_{i+1})$. Since K' has the chain property, this means $(K')^{-1}(b_{i+1}) \subset (K')^{-1}(b_i)$, i.e. $b_i \sqsubset^T_K b_{i+1}$.

We now have $b_1 \sqsubset^T_K \cdots \sqsubset^T_K b_s$; a chain of s strict inequalities in $\sqsubseteq^T_K$. The corresponding ranks $[b_1], \ldots, [b_s]$ are all distinct and lie inside $\mathcal{Y}$. But now we have found $s = |\mathcal{X}|$ distinct elements of $\mathcal{Y}$, so $|\mathcal{X}| \leq |\mathcal{Y}|$ as promised.

Repeating this argument with the roles of $\mathcal{X}$ and $\mathcal{Y}$ interchanged, we find that $|\mathcal{Y}| \leq |\mathcal{X}|$ also, and therefore $|\mathcal{X}| = |\mathcal{Y}|$.

To conclude, note that $\mathsf{ranks}(\preceq_K^T) \in \{|\mathcal{X}|, |\mathcal{X}|+1\}$, since there can exist at most one rank which was excluded from $\mathcal{X}$ (namely, those $a \in A$ with $K'(a) = \emptyset$). For identical reasons, $\mathsf{ranks}(\sqsubseteq_K^T) \in \{|\mathcal{Y}|, |\mathcal{Y}|+1\}$. Since $|\mathcal{X}| = |\mathcal{Y}|$, it is clear that $\mathsf{ranks}(\preceq_K^T)$ and $\mathsf{ranks}(\sqsubseteq_K^T)$ can differ by at most one, as required.

Now we prove the "if" statement. Let K be a tournament. We have $|\mathsf{ranks}(\preceq_K^T) - \mathsf{ranks}(\sqsubseteq_K^T)| \leq 1$, and must show there is tournament K' with the chain property such that $T(K) = (\leqslant_{K'}^{\mathcal{A}}, \leqslant_{K'}^{\mathcal{B}})$.

Let $X_1 < \cdots < X_s$ and $Y_1 < \cdots < Y_t$ be the ranks of $\preceq_K^T$ and $\sqsubseteq_K^T$ respectively. By hypothesis $|s-t| \leq 1$. Define $g : \{1, \ldots, s\} \to \{0, \ldots, t\}$ by

$$g(i) = \begin{cases} i, & s \in \{t-1, t\} \\ i-1, & s = t+1. \end{cases}$$

Not that the two cases above cover all possibilities, since $|s-t| \leq 1$. For $i \in [s]$, write

$$N_i = \bigcup_{0 \leq j \leq g(i)} Y_j,$$

where $Y_0 := \emptyset$. Note that $g(i+1) = g(i)+1$, and consequently

$$N_{i+1} = \bigcup_{j \leq g(i)+1} Y_j = N_i \cup Y_{g(i)+1} = N_i \cup Y_{g(i+1)}.$$

Since $g(i+1) > 0$ we have $Y_{g(i+1)} \neq \emptyset$, and thus $N_{i+1} \supset N_i$ for all $i < s$.

Now, for any $a \in A$, let $p(a) \in [s]$ be the unique integer such that $a \in X_{p(a)}$; such $p(a)$ always exists since $\{X_1, \ldots, X_s\}$ is a partition of A. Note that due to the assumption on the ordering of the X_i, we have $a \preceq_K^T a'$ if and only if $p(a) \leq p(a')$.

Let K' be the unique tournament such that $K'(a) = N_{p(a)}$ for each $a \in A$. Since $N_1 \subset \cdots \subset N_s$, we have

$$\begin{aligned} a \preceq_K^T a' &\iff p(a) \leq p(a') \\ &\iff N_{p(a)} \subseteq N_{p(a')} \\ &\iff K'(a) \subseteq K'(a') \\ &\iff a \leqslant_{K'}^{\mathcal{A}} a', \end{aligned} \tag{3.9}$$

i.e. $\preceq_K^T = \leqslant_{K'}^{\mathcal{A}}$. Since $\preceq_K^T$ is a total preorder, this shows that K' has the chain property.

It only remains to show that $\sqsubseteq_K^T = \leqslant_{K'}^{\mathcal{B}}$. First note that if $a \in X_i$ and $b \in Y_j$, the fact that $\{Y_1, \ldots, Y_t\}$ are disjoint implies

$$\begin{aligned} a \in (K')^{-1}(b) &\iff b \in K'(a) = N_i = \bigcup_{0 \leq k \leq g(i)} Y_k \\ &\iff j \leq g(i). \end{aligned}$$

Hence $(K')^{-1}(b)$ only depends on j: every $b \in Y_j$ shares the same neighbourhood M_j, given by

$$M_j = \bigcup_{i\in[s]:\ g(i)\geq j} X_i.$$

Note that if $1 \leq j < t$,

$$\begin{aligned} M_j &= \bigcup_{i\in[s]:\ g(i)\geq j} X_i \\ &= \left(\bigcup_{i\in[s]:\ g(i)\geq j+1} X_i\right) \cup \left(\bigcup_{i\in g^{-1}(j)} X_i\right) \\ &= M_{j+1} \cup \bigcup_{i\in g^{-1}(j)} X_i. \end{aligned}$$

Since $1 \leq j < t$ we have

$$g^{-1}(j) = \begin{cases} \{j\}, & s \in \{t-1, t\} \\ \{j+1\}, & s = t+1. \end{cases}$$

In particular $g^{-1}(j) \neq \emptyset$, which means $\bigcup_{i\in g^{-1}(j)} X_i \neq \emptyset$ and thus $M_j \supset M_{j+1}$ for all $1 \leq j < t$.

Finally, since $(K')^{-1}(b) = M_j$ for $b \in Y_j$ and $M_1 \supset \cdots \supset M_t$, an argument almost identical to (3.9) shows that $\sqsubseteq_K^T = \leqslant_{K'}^{\mathcal{B}}$.

We have shown that $T(K) = (\leqslant_{K'}^{\mathcal{A}}, \leqslant_{K'}^{\mathcal{B}})$ and that K' has the chain property, and the proof is therefore complete. $\square$

3.5.2 Interleaving Operators

According to Theorem 3.5.1, to construct a chain-definable operator it is enough to ensure that the number of ranks of $\preceq_K^T$ and $\sqsubseteq_K^T$ differ by at most one. A simple way to achieve this is to iteratively select and remove the top-ranked players of A and B simultaneously, until one of A or B is exhausted. We call such operators *interleaving operators*. Closely related ranking methods have been previously introduced for non-bipartite tournaments by Bouyssou (2004).

Formally, our procedure is defined by two functions f and g which select the next top ranks given a tournament K and subsets $A' \subseteq A$, $B' \subseteq B$ of the remaining players.

Definition 3.5.1. *An* $\mathcal{A}$-selection function *is a mapping* $f : \mathcal{K} \times 2^{\mathbb{N}} \times 2^{\mathbb{N}} \to 2^{\mathbb{N}}$ *such that for any tournament* K, $A' \subseteq A$ *and* $B' \subseteq B$:

1. $f(K, A', B') \subseteq A'$;
2. *If* $A' \neq \emptyset$ *then* $f(K, A', B') \neq \emptyset$;

3. $f(K, A', \emptyset) = A'$

Similarly, a $\mathcal{B}$-selection function *is a mapping* $g : \mathcal{K} \times 2^{\mathbb{N}} \times 2^{\mathbb{N}} \to 2^{\mathbb{N}}$ *such that*

1. $g(K, A', B') \subseteq B'$*;*

2. If $B' \neq \emptyset$ *then* $g(K, A', B') \neq \emptyset$*;*

3. $g(K, \emptyset, B') = B'$

The corresponding interleaving operator ranks players according to how soon they are selected in this way; the earlier the better.

Definition 3.5.2. *Let f and g be selection functions and K a tournament. Write $A_0 = A$, $B_0 = B$, and for $i \geq 0$:*

$$A_{i+1} = A_i \setminus f(K, A_i, B_i); \quad B_{i+1} = B_i \setminus g(K, A_i, B_i)$$

For $a \in A$ and $b \in B$, write $r(a) = \max\{i \mid a \in A_i\}$ and $s(b) = \max\{i \mid b \in B_i\}$. We define the corresponding interleaving operator $T = T^{\mathsf{int}}_{f,g}$ *by $a \preceq^T_K a'$ iff $r(a) \geq r(a')$ and $b \sqsubseteq^T_K b'$ iff $s(b) \geq s(b')$.*

Note that A_i and B_i are the players left remaining after i applications of f and g, i.e. after removing the top i ranks from both sides. Taking the maximum index in the definition of r and s is justified by the following result, which shows the interleaving process eventually terminates with $A_i = B_i = \emptyset$. Since $A_{i+1} \subseteq A_i$ and $B_{i+1} \subseteq B_i$, this shows r and s are well-defined.

Proposition 3.5.1. *Let f and g be selection functions. Fix a tournament K and let A_i, B_i $(i \geq 0)$ be as in Definition 3.5.2. Then there are $j, j' \geq 1$ such that $A_j = \emptyset$ and $B_{j'} = \emptyset$. Moreover, there is $t \geq 1$ such that both $A_t = B_t = \emptyset$.*

Proof. Suppose $i \geq 0$ and $A_i \neq \emptyset$. Then properties (1) and (2) for f in Definition 3.5.1 imply that $\emptyset \subset f(K, A_i, B_i) \subseteq A_i$, and consequently $A_{i+1} = A_i \setminus f(K, A_i, B_i) \subset A_i$.

Supposing that $A_j \neq \emptyset$ for all $j \geq 0$, we would have $A_0 \supset A_1 \supset A_2 \supset \cdots$ which clearly cannot be the case since each A_j lies inside A which is a finite set. Hence there is $j \geq 1$ such that $A_j = \emptyset$. Moreover, since $A_j \supseteq A_{j+1} \supseteq A_{j+2} \supseteq \cdots$, we have $A_k = \emptyset$ for all $k \geq j$.

An identical argument with g shows that there is $j' \geq 1$ such that $B_{j'} = \emptyset$ and $B_k = \emptyset$ for all $k \geq j'$.

Taking $t = \max\{j, j'\}$, we have $A_t = B_t = \emptyset$ as required. ☐

Before giving a concrete example of an interleaving operator, we note that interleaving is not just *one* way to satisfying **Chain-def**, it is the *only* way.

Theorem 3.5.2. *An operator T satisfies **Chain-def** if and only if $T = T^{\mathsf{int}}_{f,g}$ for some selection functions f, g.*

Theorem 3.5.2 justifies our study of interleaving operators, and provides a different perspective on chain-definability via the selection functions f and g.

Proof. Throughout the proof we will refer to a pair of total preorders $(\preceq, \sqsubseteq)$ as "chain-definable" if there is a chain tournament K such that $\preceq \, = \, \leqslant_K^{\mathcal{A}}$ and $\sqsubseteq \, = \, \leqslant_K^{\mathcal{B}}$.

First we prove the "if" direction. Let $T = T_{f,g}^{\mathsf{int}}$ be an interleaving operator with selection functions f, g, and fix a tournament K. We will show that $T(K)$ is chain-definable.

As per Proposition 3.5.1, let $j, j' \geq 1$ be the minimal integers such that $A_j = \emptyset$ and $B_{j'} = \emptyset$. Then we have $A_0 \supset \cdots \supset A_{j-1} \supset A_j = \emptyset$ and $B_0 \supset \cdots \supset B_{j'-1} \supset B_{j'} = \emptyset$.

Recall that, for $a \in A$, we have by definition $r(a) = \max\{i \mid a \in A_i\}$, which is the unique integer such that $a \in A_{r(a)} \setminus A_{r(a)+1}$. Since $a \preceq_K^T a'$ iff $r(a) \geq r(a')$, it follows that the non-empty sets $A_0 \setminus A_1, \ldots, A_{j-1} \setminus A_j$ form the ranks of the total preorder $\preceq_K^T$ (that is, the equivalence classes of the symmetric closure $\approx_K^T$). Thus, $\preceq_K^T$ has j ranks. An identical argument shows that $\sqsubseteq_K^T$ has j' ranks.

It follows from Theorem 3.5.1 that $T(K)$ is chain-definable if and only if $|j - j'| \leq 1$. If $j = j'$ this is clear. Suppose $j < j'$. Then $A_j = \emptyset$ and $B_j \neq \emptyset$. By property (3) for g in Definition 3.5.1, we have $g(K, A_j, B_j) = g(K, \emptyset, B_j) = B_j$. But this means $B_{j+1} = B_j \setminus g(K, A_j, B_j) = B_j \setminus B_j = \emptyset$. Consequently $j' = j+1$, and $|j - j'| = |-1| = 1$

If instead $j > j'$, then a similar argument using property (3) for f in Definition 3.5.1 shows that $j = j' + 1$, and we have $|j - j'| = |1| = 1$.

Hence $|j - j'| \leq 1$ in all cases, and $T(K)$ is chain-definable as required.

Now for the "only if" direction. Suppose T satisfies **Chain-def**. We will define f, g such that $T = T_{f,g}^{\mathsf{int}}$. The idea behind the construction is straightforward: since f and g pick off the next-top-ranked As and Bs at each iteration, simply define $f(K, A_i, B_i)$ as the maximal elements of A_i with respect to the existing ordering $\preceq_K^T$ (g will be defined similarly). The interleaving algorithm will then select the ranks of $\preceq_K^T$ and $\sqsubseteq_K^T$ one-by-one; the fact that $T(K)$ is chain-definable ensures that we select *all* the ranks before the iterative procedure ends. The formal details follow.

Fix a tournament K. By Theorem 3.5.1, $|\mathsf{ranks}(\preceq_K^T) - \mathsf{ranks}(\sqsubseteq_K^T)| \leq 1$. Taking $t = \max\{\mathsf{ranks}(\preceq_K^T), \mathsf{ranks}(\sqsubseteq_K^T)\}$, we can write $X_1, \ldots, X_t \subseteq A$ and $Y_1, \ldots, Y_t \subseteq B$ for the ranks of $\preceq_K^T$ and $\sqsubseteq_K^T$ respectively, possibly with $X_1 = \emptyset$ if $\mathsf{ranks}(\sqsubseteq_K^T) = 1 + \mathsf{ranks}(\preceq_K^T)$ or $Y_1 = \emptyset$ if $\mathsf{ranks}(\preceq_K^T) = 1 + \mathsf{ranks}(\sqsubseteq_K^T)$. Note that $X_i, Y_i \neq \emptyset$ for $i > 1$. Assume these sets are ordered such that $a \preceq_K^T a'$ iff $i \leq j$ whenever $a \in X_i$ and $a' \in X_j$ (and similar for the Y_i). Also note that the $X_i \cap X_j = \emptyset$ for $i \neq j$ (and similar for the Y_i).

Now set[16]

$$f(K, A', B') = \begin{cases} \max(A', \preceq_K^T), & B' \neq \emptyset \\ A', & B' = \emptyset \end{cases}$$

$$g(K, A', B') = \begin{cases} \max(B', \sqsubseteq_K^T), & A' \neq \emptyset \\ B', & A' = \emptyset \end{cases}$$

It is not difficult to see that f and g satisfy the conditions of Definition 3.5.1 for selection functions. We claim that for with A_i, B_i denoting the interleaving sets for K and f, g, for all $0 \leq i \leq t$ we have

$$A_i = \bigcup_{j=1}^{t-i} X_j, \quad B_i = \bigcup_{j=1}^{t-i} Y_j \tag{3.10}$$

For $i = 0$ this is clear: since $X_1, \ldots, X_t$ contains all ranks of $\preceq_K^T$ we have $\bigcup_{j=1}^{t-0} = X_1 \cup \cdots \cup X_t = A = A_0$ (and similar for B).

Now suppose (3.10) holds for some $0 \leq i < t$. We will show that $f(K, A_i, B_i) = X_{t-i}$ by considering three possible cases, at least one of which must hold.

Case 1: ($A_i \neq \emptyset$, $B_i \neq \emptyset$). Here we have

$$\begin{aligned} f(K, A_i, B_i) &= \max(A_i, \preceq_K^T) \\ &= \max(X_1 \cup \cdots \cup X_{t-i}, \preceq_K^T) \\ &= X_{t-i} \end{aligned}$$

since the X_j form (disjoint) ranks of $\preceq_K^T$ with $X_j \prec X_k$ for $j < k$.

Case 2: ($B_i = \emptyset$). Here we have $\bigcup_{j=1}^{t-i} Y_j = \emptyset$. Since $t - i \geq 1$ and $Y_j \neq \emptyset$ for $j > 1$, it must be the case that $t - i = 1$ and $B_i = Y_1 = \emptyset$. Consequently by the induction hypothesis we have $A_i = \bigcup_{j=1}^{1} X_j = X_1$, and thus

$$\begin{aligned} f(K, A_i, B_i) &= f(K, A_i, \emptyset) \\ &= A_i \\ &= X_1 \\ &= X_{t-i} \end{aligned}$$

Case 3: ($A_i = \emptyset$). By a similar argument as in case 2, we must have $t - i = 1$ and $A_i = X_1 = \emptyset$. Using the fact that $f(K, A_i, B_i) \subseteq A_i$ we get

$$\begin{aligned} f(K, A_i, B_i) &= \underbrace{f(K, \emptyset, B_i)}_{\subseteq \emptyset} \\ &= \emptyset \\ &= X_1 \\ &= X_{t-i} \end{aligned}$$

[16] Here $\max(Z, \preceq) = \{z \in Z \mid \nexists z' \in Z : z \prec z'\}$, for any set Z and a total preorder $\preceq$ on Z (with strict part $\prec$).

We have now covered all cases, and have shown that $f(K, A_i, B_i) = X_{t-i}$ must hold. Consequently, using again the fact that the X_j are disjoint,

$$\begin{aligned} A_{i+1} &= A_i \setminus f(K, A_i, B_i) \\ &= \left(\bigcup_{j=1}^{t-i} X_j \right) \setminus X_{t-i} \\ &= \bigcup_{j=1}^{t-(i+1)} X_j \end{aligned}$$

as required. By almost identical arguments we can show that $g(K, A_i, B_i) = Y_{t-i}$, and thus $B_{i+1} = \bigcup_{j=1}^{t-(i+1)} Y_j$ also. By induction, (3.10) holds for all $0 \leq i \leq t$.

It remains to show that $a \preceq_K^T a'$ iff $a \preceq_K^{T_{f,g}^{\mathsf{int}}} a'$ and that $b \sqsubseteq_K^T b'$ iff $b \sqsubseteq_K^{T_{f,g}^{\mathsf{int}}} b'$.

For $a \in A$, let $p(a)$ be the unique integer such that $a \in X_{p(a)}$, i.e. $p(a)$ is the index of the rank of a in the ordering $\preceq_K^T$. Note that we have

$$a \in A_i = X_1 \cup \cdots \cup X_{t-i} \iff t - i \geq p(a)$$

and therefore

$$r(a) = \max\{i \mid a \in A_i\} = \max\{i \mid t - i \geq p(a)\} = t - p(a)$$

Using the fact that $X_i \prec X_j$ for $i < j$, we get

$$\begin{aligned} a \preceq_K^{T_{f,g}^{\mathsf{int}}} a' &\iff r(a) \geq r(a') \\ &\iff t - p(a) \geq t - p(a') \\ &\iff p(a) \leq p(a') \\ &\iff a \preceq_K^T a' \end{aligned}$$

A similar argument shows that $b \sqsubseteq_K^T b'$ iff $b \sqsubseteq_K^{T_{f,g}^{\mathsf{int}}} b'$ for any $b, b' \in B$. Since K was arbitrary, we have shown that $T = T_{f,g}^{\mathsf{int}}$ as required. □

We now come to an important example.

Example 3.5.1. *Define the* cardinality-based interleaving operator $T_{\mathsf{CI}} = T_{f,g}^{\mathsf{int}}$ *where* $f(K, A', B') = \arg\max_{a \in A'} |K(a) \cap B'|$ *and* $g(K, A', B') = \arg\min_{b \in B'} |K^{-1}(b) \cap A'|$, *so that the "winners" at each iteration are the As with the most wins, and the Bs with the least losses, when restricting to A' and B' only. We take the* arg min/arg max *to be the emptyset whenever A' or B' is empty.*

Table 3.1 shows the iteration of the algorithm for a 4×5 tournament K. In each row i we show K with the rows and columns of $A \setminus A_i$ and $B \setminus B_i$ greyed out, so as to make it more clear how the f and g values are calculated.[17]

Table 3.1: Iteration of the interleaving algorithm for T_{CI}

i	K	A_i	B_i	f	g	K'_i
0	$\begin{bmatrix} 1&1&1&1&0 \\ 0&1&0&0&1 \\ 0&1&0&1&1 \\ 0&1&1&0&0 \end{bmatrix}$	$\{1,2,3,4\}$	$\{1,2,3,4,5\}$	$\{1\}$	$\{1\}$	$\begin{bmatrix} 1&1&1&1&\color{red}\mathbf{1} \\ 0&1&0&0&1 \\ 0&1&0&1&1 \\ 0&1&1&0&0 \end{bmatrix}$
1	$\begin{bmatrix} \color{gray}1&\color{gray}1&\color{gray}1&\color{gray}1&\color{gray}0 \\ \color{gray}0&1&0&0&1 \\ \color{gray}0&1&0&1&1 \\ \color{gray}0&1&1&0&0 \end{bmatrix}$	$\{2,3,4\}$	$\{2,3,4,5\}$	$\{3\}$	$\{3,4\}$	$\begin{bmatrix} 1&1&1&1&\color{red}\mathbf{1} \\ 0&1&0&0&1 \\ 0&1&\color{red}\mathbf{1}&1&1 \\ 0&1&\color{red}\mathbf{0}&0&0 \end{bmatrix}$
2	$\begin{bmatrix} \color{gray}1&\color{gray}1&\color{gray}1&\color{gray}1&\color{gray}0 \\ \color{gray}0&1&\color{gray}0&\color{gray}0&1 \\ \color{gray}0&\color{gray}1&\color{gray}0&\color{gray}1&\color{gray}1 \\ \color{gray}0&1&\color{gray}1&\color{gray}0&0 \end{bmatrix}$	$\{2,4\}$	$\{2,5\}$	$\{2\}$	$\{5\}$	-
3	$\begin{bmatrix} \color{gray}1&\color{gray}1&\color{gray}1&\color{gray}1&\color{gray}0 \\ \color{gray}0&\color{gray}1&\color{gray}0&\color{gray}0&\color{gray}1 \\ \color{gray}0&\color{gray}1&\color{gray}0&\color{gray}1&\color{gray}1 \\ \color{gray}0&1&\color{gray}1&\color{gray}0&\color{gray}0 \end{bmatrix}$	$\{4\}$	$\{2\}$	$\{4\}$	$\{2\}$	-
4	-	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	-

For brevity we also write f and g in place of $f(K, A_i, B_i)$ and $g(K, A_i, B_i)$ respectively.

The r and s values can be read off as 0, 2, 1, 3 for A and 0, 3, 1, 1, 2 for B, giving the ranking on A as $4 \prec 2 \prec 3 \prec 1$, and the ranking on B as $2 \sqsubset 5 \sqsubset 3 \approx 4 \sqsubset 1$. Note also that each $f(K, A_i, B_i)$ is a rank of $\preceq_K^T$ (and similar for $g(K, A_i, B_i)$), so the rankings can in fact be read off by looking at the f and g columns of Table 3.1.

The interleaving algorithm can also be seen as a greedy algorithm for converting K into a chain graph directly. Indeed, by setting the neighbourhood of each $a \in f(K, A_i, B_i)$ to B_i, and removing each $b \in g(K, A_i, B_i)$ from the neighbourhoods of all $a \in A_{i+1}$, we eventually obtain a chain graph. We show this process in the K'_i column of Table 3.1, where only three entries need to be changed.[18] The selection functions f and g can therefore be seen as *heuristics* with the goal of finding a chain graph "close" to K.

The operator T_{CI} from Example 3.5.1 uses simple cardinality-based heuristics, and can be seen as a chain-definable version of T_{count} (which is not chain-definable). It is also the bipartite counterpart to repeated applications of Copeland's rule (Bouyssou 2004). Note that $f(K, A_i, B_i)$ and $g(K, A_i, B_i)$ can be computed in $O(N^2)$ time at each iteration i, where $N = |A| + |B|$. Since there cannot be more than N iterations, it follows that the rankings of T_{CI} can be computed in $O(N^3)$ time.

3.5.3 Axiom Compatibility

We now revisit the axioms of Section 3.3 in relation to chain-definable operators in general and T_{CI} specifically. Firstly, the weakening of **Chain-min** pays off: **Chain-def** is compatible with all our axioms.

[17] Note that while f and g for T_{CI} are independent of the greyed out entries, we do not require this property for selection functions in general.

[18] In this example $\mathcal{M}(K)$ contains a single tournament a distance of 2 from K, so T_{CI} makes one more change than necessary.

Theorem 3.5.3. *For each of* ***Anon****,* ***Dual****,* ***IIM****,* ***Mon*** *and* ***Pos-resp****, there exists an operator satisfying* ***Chain-def*** *and the stated property.*

Proof. Since **Chain-min** implies **Chain-def**, Theorem 3.3.2 implies the existence of an operator with **Chain-def** and **Dual**, and an operator with **Chain-def** and **Mon**. Moreover, the trivial operator which ranks all As and Bs equally satisfies **Chain-def**, **Anon** and **IIM**. It only remains to show that there is an operator satisfying both **Chain-def** and **Pos-resp**.

To that end, for any tournament K, define K' by

$$K'_{ab} = \begin{cases} 1, & b \leq |K(a)| \\ 0, & b > |K(a)| \end{cases}$$

Note that $K'(a) = \{1, \ldots, |K(a)|\}$ for $|K(a)| > 0$. Consequently $K'(a) \subseteq K'(a')$ iff $|K(a)| \leq |K(a)|$. We see that K' has the chain property, and the operator T defined by $T(K) = (\leqslant^{\mathcal{A}}_{K'}, \leqslant^{\mathcal{B}}_{K'})$ satisfies **Chain-def**. In particular, $a \preceq^T_K a'$ iff $|K(a)| \leq |K(a')|$.

To show **Pos-resp**, suppose $a \preceq^T_K a'$ and $K_{a',b} = 0$ for some $a, a' \in A$ and $b \in B$. Write $\hat{K} = K + \mathbf{1}_{a',b}$.

Since $a \preceq^T_K a'$ implies $|K(a)| \leq |K(a')|$, we have $|\hat{K}(a')| = 1 + |K(a')| > |K(a)| = |\hat{K}(a)|$, and therefore $a \prec^T_{\hat{K}} a'$ as required for **Pos-resp**. $\square$

Unfortunately, these cannot all hold at the same time. Indeed, taking $K = \left[\begin{smallmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{smallmatrix}\right]^\top$ and assuming **Anon** and **Pos-resp**, the ranking on A is fully determined as $1 \prec 2 \approx 3 \prec 4$, and $\mathsf{ranks}(\preceq^T_K) = 3$. However, **Anon** with **Dual** implies the ranking of B is flat, i.e. $\mathsf{ranks}(\sqsubseteq^T_K) = 1$. This contradicts **Chain-def** by Theorem 3.5.1, yielding the following impossibility result.

Theorem 3.5.4. *There is no operator satisfying* ***Chain-def****,* ***Anon****,* ***Dual*** *and* ***Pos-resp****.*

Proof. For contradiction, suppose there is an operator T satisfying the stated axioms. Consider

$$K = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$

and two tournaments obtained by removing a single 1 entry:

$$K_1 = \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{0} \\ 1 & 0 \\ 1 & 1 \end{bmatrix}, \quad K_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & \mathbf{0} \end{bmatrix}.$$

Now, **Anon** in K_1 gives $1 \approx^T_{K_1} 2$ (e.g. take $\sigma = (1\ 2)$, $\pi = \mathrm{id}_B$). In particular, $1 \preceq^T_{K_1} 2$, so **Pos-resp** implies $1 \prec^T_K 2$. A similar argument with K_2 shows that $3 \approx^T_{K_2} 4$ and $3 \prec^T_K 4$.

On the other hand, applying **Anon** to K directly with $\sigma = (2\ 3)$ and $\pi = (1\ 2)$, we see that $2 \approx^T_K 3$. The ranking of A is thus fully determined as $1 \prec 2 \approx 3 \prec 4$. In particular, $\mathsf{ranks}(\preceq^T_K) = 3$.

But now considering the dual tournament $\overline{K} = \left[\begin{smallmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{smallmatrix}\right]$ and applying permutations $\sigma = (1\ 2)$ and $\pi = (2\ 3)$, we obtain $1 \approx^{T}_{\overline{K}} 2$ by **Anon**, i.e. the A ranking in $\overline{K}$ is flat. By **Dual** this implies the B ranking in K is flat, i.e. $\mathsf{ranks}(\sqsubseteq^T_K) = 1$. We see that $\mathsf{ranks}(\preceq^T_K)$ and $\mathsf{ranks}(\sqsubseteq^T_K)$ differ by 2, contradicting **Chain-def** according to Theorem 3.5.1. $\square$

For interleaving operators, we have the following sufficient conditions for $T^{\mathsf{int}}_{f,g}$ to satisfy various axioms.

Lemma 3.5.1. *Let $T = T^{int}_{f,g}$ be an interleaving operator.*

1. *If for any tournament K, $A' \subseteq A$, $B' \subseteq B$ and for any pair of permutations $\sigma : A \to A$ and $\pi : B \to B$ we have*

$$f(\pi(\sigma(K)), \sigma(A'), \pi(B')) = \sigma(f(K, A', B'))$$
$$g(\pi(\sigma(K)), \sigma(A'), \pi(B')) = \pi(g(K, A', B'))$$

   *then T satisfies **Anon**.*

2. *If for any tournament K and $A' \subseteq A$, $B \subseteq B$ we have*

$$g(K, A', B') = f(\overline{K}, B', A')$$

   *then T satisfies **Dual**.*

3. *If for any tournament K, $A' \subseteq A$, $B' \subseteq B$ and $a, a' \in A'$ we have*

$$K(a) \subseteq K(a') \implies a \notin f(K, A', B') \text{ or } a' \in f(K, A', B')$$

   *then T satisfies **Mon**.*

Proof. We take each statement in turn.

1. Let K be a tournament. For brevity, write $K' = \pi(\sigma(K))$. Let us write A_i, B_i and A'_i, B'_i ($i \geq 0$) for the sets defined in Definition 3.5.2 for K and K' respectively. We claim that for all $i \geq 0$:

$$A'_i = \sigma(A_i), \quad B'_i = \pi(B_i) \tag{3.11}$$

   For $i = 0$ this is trivial since $A'_0 = A = \sigma(A) = \sigma(A_0)$ since σ is a bijection. The fact that $B'_0 = \pi(B_0)$ is shown similarly.

   Suppose that (3.11) holds for some $i \geq 0$. Then applying our assumption on f:

$$\begin{aligned} A'_{i+1} &= A'_i \setminus f(K', A'_i, B'_i) \\ &= \sigma(A_i) \setminus f(K', \sigma(A_i), \pi(B_i)) \\ &= \sigma(A_i) \setminus \sigma(f(K, A_i, B_i)) \\ &= \sigma(A_i \setminus f(K, A_i, B_i)) \\ &= \sigma(A_{i+1}) \end{aligned}$$

(note that $\sigma(X) \setminus \sigma(Y) = \sigma(X \setminus Y)$ holds for any sets X, Y due to injectivity of σ). Using the assumption on g we can show that $B'_{i+1} = \pi(B_{i+1})$ in a similar manner. Therefore, by induction, (3.11) holds for all $i \geq 0$. This means that for any $a \in A$ we have

$$\sigma(a) \in A'_i \iff \sigma(a) \in \sigma(A_i) \iff a \in A_i$$

and therefore, with r_K and $r_{K'}$ denoting the functions $A \to \mathbb{N}_0$ defined in Definition 3.5.2 for K and K' respectively,

$$\begin{aligned} r_{K'}(\sigma(a)) &= \max\{i \mid \sigma(a) \in A'_i\} \\ &= \max\{i \mid a \in A_i\} \\ &= r_K(a) \end{aligned}$$

From this it easily follows that $a \preceq^T_K a'$ iff $\sigma(a) \preceq^T_{K'} \sigma(a')$, i.e. T satisfies **Anon**.

2. Once again, fix a tournament K and let A_i, B_i and A'_i, B'_i denote the sets from Definition 3.5.2 for K and $\overline{K}$ respectively. It is easy to show by induction that the assumption on f and g implies $A'_i = B_i$ and $B'_i = A_i$ for all $i \geq 0$. This means that for any $b \in B_K$:

$$\begin{aligned} s_K(b) &= \max\{i \mid b \in B_i\} \\ &= \max\{i \mid b \in A'_i\} \\ &= r_{\overline{K}}(b) \end{aligned}$$

which implies $b \sqsubseteq^T_K b'$ iff $b \preceq^T_{\overline{K}} b'$, as required for **Dual**.

3. Let K be a tournament and $a, a' \in A$ such that $K(a) \subseteq K(a')$. We must show that $a \preceq^T_K a'$.

   Suppose otherwise, i.e. $a' \prec^T_K a$. Then $r(a') > r(a)$. Note that by definition of r, we have $a \in A_{r(a)} \setminus A_{r(a)+1} = f(K, A_{r(a)}, B_{r(a)})$. Since $r(a') \geq r(a) + 1$ and $A_{r(a)} \supseteq A_{r(a)+1} \supseteq A_{r(a)+2} \supseteq \cdots$, we get $a' \in A_{r(a)+1} \subseteq A_{r(a)}$. In particular, $a' \notin f(K, A_{r(a)}, B_{r(a)})$.

   Piecing this all together, we have $a, a' \in A_{r(a)}$, $K(a) \subseteq K(a')$, $a \in f(K, A_{r(a)}, B_{r(a)})$ and $a' \notin f(K, A_{r(a)}, B_{r(a)})$. But this directly contradicts our assumption on f, so we are done.

□

For the specific operator T_{CI}, Lemma 3.5.1 yields the following.

Theorem 3.5.5. *T_{CI} satisfies **Chain-def**, **Anon**, **Dual** and **Mon**, and does not satisfy **IIM** or **Pos-resp**.*

Proof. We take each axiom in turn. Let f and g be the selection functions corresponding to T_{CI} from Example 3.5.1.

Chain-def. Since T_{CI} is an interleaving operator, **Chain-def** follows from Theorem 3.5.2.

Anon. Let K be a tournament and let $\sigma : A \to A$ and $\pi : B \to B$ be bijective mappings. Write $K' = \pi(\sigma(K))$. We will show that the conditions on f and g in Lemma 3.5.1 part (1) are satisfied.

Let $A' \subseteq A$ and $B' \subseteq B$. We have

$$\begin{aligned} f(K', \sigma(A'), \pi(B')) &= \underset{\hat{a} \in \sigma(A')}{\arg\max} |K'(\hat{a}) \cap \pi(B')| \\ &= \sigma(\underset{a \in A'}{\arg\max} |K'(\sigma(a)) \cap \pi(B')|) \end{aligned}$$

where we make the "substitution" $a = \sigma^{-1}(\hat{a})$. Using the definition of $K' = \pi(\sigma(K))$ it is easily seen that $K'(\sigma(a)) = \pi(K(a))$. Also, since π is a bijection we have $\pi(X) \cap \pi(Y) = \pi(X \cap Y)$ for any sets X and Y, and $|\pi(X)| = |X|$. Thus

$$\begin{aligned} f(K', \sigma(A'), \pi(B')) &= \sigma(\underset{a \in A'}{\arg\max} |K'(\sigma(a)) \cap \pi(B')|) \\ &= \sigma(\underset{a \in A'}{\arg\max} |\pi(K(a)) \cap \pi(B')|) \\ &= \sigma(\underset{a \in A'}{\arg\max} |\pi(K(a) \cap B')|) \\ &= \sigma(\underset{a \in A'}{\arg\max} |K(a) \cap B'|) \\ &= \sigma(f(K, A', B')) \end{aligned}$$

as required. The result for g follows by a near-identical argument. Thus T_{CI} satisfies **Anon** by Lemma 3.5.1 part (1).

Dual. Fix a tournament K and let $A' \subseteq A$, $B' \subseteq B$. Note that for $b \in B'$ we have

$$\begin{aligned} |K^{-1}(b) \cap A'| &= |(A \setminus \overline{K}(b)) \cap A'| \\ &= |A' \setminus \overline{K}(b)| \\ &= |A'| - |\overline{K}(b) \cap A'| \end{aligned}$$

Consequently

$$\begin{aligned} g(K, A', B') &= \underset{b \in B'}{\arg\min} |K^{-1}(b) \cap A'| \\ &= \underset{b \in B'}{\arg\min} \left(|A'| - |\overline{K}(b) \cap A'|\right) \\ &= \underset{b \in B'}{\arg\max} |\overline{K}(b) \cap A'| \\ &= f(\overline{K}, B', A') \end{aligned}$$

and, by Lemma 3.5.1 part (2), T_{CI} satisfies **Dual**.

Mon. Once again, we use Lemma 3.5.1. Let K be a tournament and $A' \subseteq A$, $B' \subseteq B$. Suppose $a, a' \in A'$ with $K(a) \subseteq K(a')$. We need to show that either $a \notin f(K, A', B')$ or $a' \in f(K, A', B')$

Suppose $a \in f(K, A', B')$. Then $a \in \arg\max_{\hat{a} \in A'} |K(\hat{a}) \cap B'|$, so $|K(a) \cap B'| \geq |K(a') \cap B'|$. On the other hand $K(a) \cap B' \subseteq K(a') \cap B'$, so $|K(a) \cap B'| \leq |K(a') \cap B'|$. Consequently $|K(a) \cap B'| = |K(a') \cap B'|$, and so $a' \in f(K, A', B')$. This shows the property required by Lemma 3.5.1 part (3) is satisfied, and thus T_{CI} satisfies **Mon**.

Pos-resp. We have show that T_{CI} satisfies **Chain-def**, **Anon** and **Dual**; due to impossibility result of Theorem 3.5.4, T_{CI} cannot satisfy **Pos-resp**.

IIM. Write

$$K_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \quad K_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Note that the first and second rows of each tournament are identical, so **IIM** would imply $1 \preceq_{K_1}^{T_{\mathsf{CI}}} 2$ iff $1 \preceq_{K_2}^{T_{\mathsf{CI}}} 2$. However, it is easily verified that $1 \prec_{K_1}^{T_{\mathsf{CI}}} 2$ whereas $2 \prec_{K_2}^{T_{\mathsf{CI}}} 1$. Therefore T_{CI} does not satisfy **IIM**. $\square$

Note that **Anon** *is* satisfied. This makes T_{CI} an important example of a well-motivated, tractable, chain-definable and anonymous operator, meeting the criteria outlined at the start of this section.

3.5.4 Axiomatic Characterisation of T_{CI}

In Theorem 3.5.5 we saw which of the axioms from Section 3.3 hold for T_{CI}. We now characterise T_{CI} axiomatically by introducing two new axioms. The first, which we call **Rank-removal**, is a technical axiom obtained via a related property specific to interleaving operators. The second, called **Argmax**, says that the maximum rank in $\preceq_K^T$ should coincide with that of $\preceq_K^{T_{\mathsf{CI}}}$. Together with **Dual** and **Chain-def**, these will characterise T_{CI}.

Unlike the axioms introduced so far, which were straightforward, general properties for ranking methods, the new axioms are geared specifically towards characterising T_{CI}. Thus, this section takes a *descriptive* perspective as opposed to the *normative* perspective of Section 3.3.

Towards the characterisation result, we first note that T_{CI} satisfies a kind of independence property for interleaving operators: $f(K, A', B')$ and $g(K, A', B')$ only depends on the sub-matrix of K with rows and columns corresponding to A' and B'. In graphical terms, the greyed out rows and columns in Table 3.1 do not affect the output of f and g. In general this does not hold for interleaving operators. We express this formally as an axiom for interleaving operators $T_{f,g}^{\mathsf{int}}$, called *sub-matrix independence*.

SMI. Let K be a tournament and A_i, B_i a pair of non-empty sets arising in the interleaving algorithm for f, g and K in Definition 3.5.2. Write $A_i = \{a_1, \ldots, a_{m'}\}$ and $B_i = \{b_1, \ldots, b_{n'}\}$, ordered such that $a_p < a_{p+1}$ and $b_q < b_{q+1}$. Let K^- be the corresponding $m' \times n'$ sub-matrix of K, where $K^-_{pq} = K_{a_p,b_q}$. Then for all $a_p \in A_i$ and $b_q \in B_i$,

$$\begin{aligned} a_p \in f(K, A_i, B_i) &\iff p \in f(K^-, [m'], [n']), \\ b_q \in g(K, A_i, B_i) &\iff q \in g(K^-, [m'], [n']). \end{aligned}$$

Note that **SMI** is a property of the selection functions f and g. In principle, it is possible that an interleaving operator T admits two pairs of selection functions (f, g) and (f', g') such that **SMI** holds for one pair but not the other. However, it will follow from later results (Proposition 3.5.4) that this is not possible: **SMI** either holds for T or does not, independently of the choice of selection functions f and g. Nevertheless, to avoid circularity we will say T satisfies **SMI** if *there exist* f, g with the **SMI** property such that $T = T^{\mathsf{int}}_{f,g}$.

First, T_{CI} does indeed satisfy **SMI**.

Proposition 3.5.2. *T_{CI} satisfies **SMI**.*

Proof. Let f and g denote the selection functions for T_{CI}. Let A_i, B_i and K^- be as in the statement of **SMI**. Note that for any $a_p \in A_i$,

$$\begin{aligned} K(a_p) \cap B_i &= \{b \in B_i \mid K_{a_p,b} = 1\} \\ &= \{b_q \mid q \in [n'], K_{a_p,b_q} = 1\} \\ &= \{b_q \mid q \in K^-(p)\}, \end{aligned}$$

so $|K(a_p) \cap B_i| = |K^-(p)| = |K^-(p) \cap [n']|$. Consequently,

$$\begin{aligned} a_p \in f(K, A_i, B_i) &\iff a_p \in \arg\max_{a \in A_i} |K(a) \cap B_i| \\ &\iff p \in \arg\max_{p' \in [m']} |K(a_{p'}) \cap B_i| \\ &\iff p \in \arg\max_{p' \in [m']} |K^-(p)| \\ &\iff p \in f(K^-, [m'], [n']) \end{aligned}$$

as required. An identical argument shows the desired property for g. Hence T_{CI} satisfies **SMI**. □

Note that in the statement of **SMI**, $[m'] = A_{K^-}$ and $[n'] = B_{K^-}$. Consequently, **SMI** implies that the ranks of A and B are fully determined by the *maximal* ranks of successively smaller sub-tournaments. This is expressed in the following result, which shows that two **SMI** operators agreeing on maximal ranks for all K must in fact be equal.

Proposition 3.5.3. *Let T and T' be interleaving operators satisfying **SMI**. Suppose that for all tournaments K,*

$$\max(A, \preceq_K^T) = \max(A, \preceq_K^{T'})$$
$$\max(B, \sqsubseteq_K^T) = \max(B, \sqsubseteq_K^{T'}).$$

Then $T = T'$.

Proof. Let f, g and f', g' be selection functions corresponding to T and T' respectively. Take a tournament K. To show $T(K) = T(K')$ it is sufficient to show $A_i = A'_i$ and $B_i = B'_i$ for all $i \geq 0$, where A_i, B_i and A'_i, B'_i are the interleaving sets from Definition 3.5.2 for T and T' respectively.

We proceed by induction on i. For $i = 0$ this is clear, since $A_0 = A'_0 = A$ and $B_0 = B'_0 = B$ by definition. Suppose $A_i = A'_i$ and $B_i = B'_i$. If $A_i = A'_i = \emptyset$ then $A_{i+1} = A'_{i+1} = \emptyset$ (since $A_{i+1} \subseteq A_i$), and similarly $B_i = B'_i = \emptyset$ implies $B_{i+1} = B'_{i+1} = \emptyset$. Hence we may assume without loss of generality that $A_i, B_i \neq \emptyset$.

Write $A_i = A'_i = \{a_1, \ldots, a_{m'}\}$ and $B_i = B'_i = \{b_1, \ldots, b_{n'}\}$, with $a_p < a_{p+1}$ and $b_q < b_{q+1}$. Let K^- be the associated sub-matrix, as in the statement of **SMI**. From property (1) from Definition 3.5.1 for f and **SMI**, we have

$$\begin{aligned} f(K, A_i, B_i) &= \{a_p \mid p \in f(K^-, [m'], [n'])\} \\ f'(K, A'_i, B'_i) &= \{a_p \mid p \in f'(K^-, [m'], [n'])\}. \end{aligned} \tag{3.12}$$

But $[m']$ and $[n']$ are the full set of players in K^-, so

$$\begin{aligned} f(K^-, [m'], [n']) &= \max(A_{K^-}, \preceq_{K^-}^T) \\ &= \max(A_{K^-}, \preceq_{K^-}^{T'}) \\ &= f'(K^-, [m'], [n']), \end{aligned}$$

where our assumption on the maximal ranks for T and T' is employed in the second step. Consulting (3.12) we see $f(K, A_i, B_i) = f'(K, A'_i, B'_i)$. An identical argument shows $g(K, A_i, B_i) = g'(K, A'_i, B'_i)$. Together with the induction hypothesis, we get

$$A_{i+1} = A_i \setminus f(K, A_i, B_i) = A'_i \setminus f'(K, A'_i, B'_i) = A'_{i+1}$$

and similarly, $B_{i+1} = B'_{i+1}$. By induction, $A_i = A'_i$ and $B_i = B'_i$ for all $i \geq 0$, and we are done. ☐

This result simplifies the task of characterising T_{CI} among the interleaving operators with **SMI**, since we only need to consider the maximal ranks of A and B. In fact, given that T_{CI} satisfies **Dual** we only need to consider the A ranking. The following axiom says that maximally-ranked players in A are exactly those for whom $|K(a)|$ is maximal; this will clearly capture the maximal ranks of T_{CI} together with **Dual**.

Argmax. $\max(A, \preceq^T_K) = \arg\max_{a\in A} |K(a)|$.

Note that **Argmax** does *not* require $\preceq^T_K$ to reduce to T_{count}, since we only consider a such that $|K(a)|$ is *maximal*. Now, since **Chain-def** characterises interleaving operators, to obtain a characterisation of T_{CI} among *all* operators it suffices to find an alternative version of **SMI** which can be applied to any operator. This is the role of the following axiom, which says that removing the maximally ranked players from each side preserves the ordering among the rest of A and B.

Rank-removal. Suppose $\max(A, \preceq^T_K) \neq A$ and $\max(B, \sqsubseteq^T_K) \neq B$. Write $A \setminus \max(A, \preceq^T_K) = \{a_1, \ldots, a_{m'}\}$ and $B \setminus \max(B, \sqsubseteq^T_K) = \{b_1, \ldots, b_{m'}\}$, ordered such that $a_p < a_{p+1}$ and $b_q < b_{q+1}$. Let K^- be the corresponding $m' \times n'$ sub-matrix of K. Then for all p, p' and q, q',

$$a_p \preceq^T_K a_{p'} \iff p \preceq^T_{K^-} p'$$
$$b_q \sqsubseteq^T_K b_{q'} \iff q \sqsubseteq^T_{K^-} q'.$$

In what sense does **Rank-removal** capture **SMI**? In the following we show it is *equivalent* to **SMI**, when taken with **Chain-def**. We need a preliminary lemma.

Lemma 3.5.2. *Let f, g be selection functions, K a tournament, and A_i, B_i sets arising in the interleaving algorithm for f, g and K. Suppose $A_i \neq \emptyset$ and $B_i \neq \emptyset$. Then*

$$f(K, A_i, B_i) = \max(A_i, \preceq^{T^{\mathsf{int}}_{f,g}}_K)$$
$$g(K, A_i, B_i) = \max(B_i, \sqsubseteq^{T^{\mathsf{int}}_{f,g}}_K).$$

Proof. We show the first statement; the second follows by an identical argument. For brevity, write T for $T^{\mathsf{int}}_{f,g}$. For the left-to-right inclusion, suppose $a \in f(K, A_i, B_i)$. Then $a \in A_i$. Take any $a' \in A_i$. We need to show $a' \preceq^T_K a$. Indeed, we have $a \in A_i \cap f(K, A_i, B_i)$, so $a \notin A_{i+1}$. Consequently $r(a) = \max\{j \mid a \in A_j\} = i$. Since $a' \in A_i$, $r(a') \geq i = r(a)$. Hence $a' \preceq^T_K a$.

For the right-to-left we show the contrapositive. Suppose $a \notin f(K, A_i, B_i)$. If $a \notin A_i$ then clearly $a \notin \max(A_i, \preceq^T_K)$. So suppose $a \in A_i$. Then $a \in A_i \setminus f(K, A_i, B_i) = A_{i+1}$. Hence $r(a) \geq i + 1 > i$. On the other hand, since $A_i \neq \emptyset$ we have by properties of the selection function f that $f(K, A_i, B_i) \neq \emptyset$. Thus there is some $a' \in A_i \cap f(K, A_i, B_i)$. We see that $r(a') = i < r(a)$, so $a \prec^T_K a'$. Hence $a \notin \max(A_i, \preceq^T_K)$, as required. □

Proposition 3.5.4. *T satisfies **Chain-def** and **Rank-removal** if and only if T is an interleaving operator satisfying **SMI**.*

Proof. For the "if" direction, suppose $T = T^{\mathsf{int}}_{f,g}$ is an interleaving operator satisfying **SMI**. Then T satisfies **Chain-def** by Theorem 3.5.2. We show **Rank-removal**. Let $\{a_1, \ldots, a_{m'}\}$, $\{b_1, \ldots, b_{n'}\}$ and K^- be as in the statement of **Rank-removal**.

Let A_i, B_i denote the interleaving sets for T applied to K, and A_i^-, B_i^- those for T applied to K^-. First we claim that for all $p \in [m']$, $q \in [n']$ and $i \geq 0$,

$$\begin{aligned} p \in A_i^- &\iff a_p \in A_{i+1}, \\ q \in B_i^- &\iff b_q \in B_{i+1}. \end{aligned} \tag{3.13}$$

We prove (3.13) by induction on i. Take $i = 0$. By definition, $A_0^- = A_{K^-} = [m']$, so $p \in A_0^-$ always holds. Recall that $\{a_1, \ldots, a_{m'}\} = A \setminus \max(A, \preceq_K^T)$. It is easily seen that $\max(A, \preceq_K^T) = f(K, A_0, B_0)$, so in fact $\{a_1, \ldots, a_{m'}\} = A_1$. Hence $a_p \in A_1$ always holds. The argument for the Bs is identical. This proves (3.13) for $i = 0$.

For the inductive step, suppose (3.13) holds for some $i \geq 0$. We consider three cases. First, suppose $A_i^- = \emptyset$. Then $A_{i+1}^- \subseteq A_i^-$ means $A_{i+1}^- = \emptyset$. On the other hand the inductive hypothesis gives $A_{i+1} = \emptyset$, so we get $A_{i+2} = \emptyset$. In particular, the first part of (3.13) holds for $i + 1$. For the second part, using property (3) from Definition 3.5.1 for the selection function g we find

$$B_{i+1}^- = B_i^- \setminus g(K^-, \emptyset, B_i^-) = B_i^- \setminus B_i^- = \emptyset.$$

By the inductive hypothesis again we have $A_{i+1} = \emptyset$, so the same line of reasoning gives $B_{i+1} = \emptyset$. Thus (3.13) holds.

The case $B_i^- = \emptyset$ is identical, using properties of f instead of g.

Now suppose both A_i^- and B_i^- are non-empty. Recall $A_i^- \subseteq A_0^- = [m']$ and $B_i^- \subseteq B_0^- = [n']$. Write $A_i^- = \{p_1, \ldots, p_{m''}\}$ and $B_i^- = \{q_1, \ldots, q_{n''}\}$ in increasing order. Let K^{--} be the $m'' \times n''$ sub-matrix of K^- formed by A_i^- and B_i^-, i.e. $K_{st}^{--} = K_{p_s,q_t}^-$. Since $T = T^{\mathsf{int}}_{f,g}$ satisfies **SMI**, we get that for all $s \in [m'']$ and $t \in [n'']$,

$$p_s \in f(K^-, A^-i, B_i^-) \iff s \in f(K^{--}, [m''], [n'']).$$

Now, recall that $K_{pq}^- = K_{a_p,b_q}$. Hence

$$K_{st}^{--} = K_{p_s,q_t}^- = K_{a_{p_s},b_{q_t}}.$$

By the inductive hypothesis,

$$\begin{aligned} A_{i+1} &= \{a_p \mid p \in A_i^-\} = \{a_{p_s} \mid s \in [m'']\}, \\ B_{i+1} &= \{b_q \mid q \in B_i^-\} = \{b_{q_t} \mid t \in [n'']\}. \end{aligned}$$

We see that K^{--} can also be viewed as the $m'' \times n''$ sub-matrix of K formed by A_{i+1} and B_{i+1}. Applying **SMI** in this instance, we find

$$a_{p_s} \in f(K, A_{i+1}, B_{i+1}) \iff s \in f(K^{--}, [m''], [n'']).$$

Putting things together,

$$a_{p_s} \in f(K, A_{i+1}, B_{i+1}) \iff p_s \in f(K^-, A_i^-, B_i^-).$$

Consequently, for any $p \in [m']$ we get

$$\begin{aligned} p \in A_{i+1}^- &\iff p \in A_i^- \text{ and } p \notin f(K^-, A_i^-, B_i^-) \\ &\iff \exists s \in [m''] : \ p = p_s \text{ and } p_s \notin f(K^-, A_i^-, B_i^-) \\ &\iff \exists s \in [m''] : \ p = p_s \text{ and } a_{p_s} \notin f(K, A_{i+1}, B_{i+1}) \\ &\iff a_p \in A_{i+1} \text{ and } a_p \notin f(K, A_{i+1}, B_{i+1}) \\ &\iff a_p \in A_{i+2}. \end{aligned}$$

An identical argument shows $q \in B^-i+1 \iff b_q \in B_{i+2}$. By induction, this completes the proof of (3.13).

Finally, for any $p \in [m']$ we get

$$\begin{aligned} r_{K^-}(p) &= \max\{i \mid p \in A_i^-\} \\ &= \max\{i \mid a_p \in A_{i+1}\} \\ &= \max\{i-1 \mid a_p \in A_i\} \\ &= -1 + r_K(a_p), \end{aligned}$$

where we use (3.13) and the fact that $a_p \in A_1$. Since the additive constant of -1 does not affect the ranking, we see that the ranking of $[m']$ in $\preceq^T_{K^-}$ corresponds exactly to that of $\{a_1, \ldots, a_{m'}\}$ in $\preceq^T_K$, as required for **Rank-removal**. The case for the B ranking is identical, and we are done.

Now for the "only if" direction, suppose T satisfies **Chain-def** and **Rank-removal**. By Theorem 3.5.2 there are selection functions f and g such that $T = T^{\mathsf{int}}_{f,g}$. We need to show that **SMI** holds. Let K, $A_i = \{a_1, \ldots, a_{m'}\}, B_i = \{b_1, \ldots, b_{n'}\}$ and K^- be as in the statement of **SMI**. Without loss of generality, $i > 0$. A simple induction shows that

$$A_j = A \setminus \bigcup_{k<j} f(K, A_k, B_k)$$

for all $j \geq 0$. From Lemma 3.5.2 we see that A_i is the result of removing the top $i-1$ ranks of players in A, according to the ranking $\preceq^T_K$. Applying **Rank-removal** $i-1$ times, we get

$$a_p \preceq^T_K a_{p'} \iff p \preceq^T_{K^-} p'$$

for all $p, p' \in [m']$. Consequently, using Lemma 3.5.2 again,

$$\begin{aligned} a_p \in f(K, A_i, B_i) &\iff a_p \in \max(A_i, \preceq^T_K) \\ &\iff \forall p' \in [m'] : \ a_{p'} \preceq^T_K a_p \\ &\iff \forall p' \in [m'] : \ p' \preceq^T_{K^-} p \\ &\iff p \in \max([m'], \preceq^T_{K^-}) \\ &\iff p \in f(K^-, [m'], [n']) \end{aligned}$$

as required for **SMI**. One can show $b_q \in g(K, A_i, B_i) \iff q \in g(K^-, [m'], [n'])$ by an identical argument. $\square$

Finally, we can state the axiomatic characterisation of T_{CI}.

Theorem 3.5.6. *T_{CI} is the unique operator satisfying **Dual**, **Chain-def**, **Rank-removal** and **Argmax**.*

Proof. We have already seen in Theorem 3.5.5 that T_{CI} satisfies **Dual** and **Chain-def**. For **Argmax**, note that for any tournament K,

$$\begin{aligned}\max(A, \preceq_K^{T_{\mathsf{CI}}}) &= f(K, A, B) \\ &= \operatorname*{arg\,max}_{a \in A} |K(a) \cap B| \\ &= \operatorname*{arg\,max}_{a \in A} |K(a)|\end{aligned}$$

directly from the definition of the selection function f for T_{CI}. Finally, **Rank-removal** follows from Proposition 3.5.4 since T_{CI} is an interleaving operator with **SMI** (by Proposition 3.5.2).

For uniqueness, suppose some operator T also satisfies the stated axioms. By Proposition 3.5.4, T is an interleaving operator satisfying **SMI**. To show $T = T_{\mathsf{CI}}$ it is sufficient by Proposition 3.5.3 to show that T and T_{CI} agree on maximal ranks for all tournaments K. Clearly this is the case for the ranking of A, since **Argmax** completely prescribes $\max(A, \preceq_K^T)$ solely in terms of K. Moreover, by **Dual** we have $\sqsubseteq_K^T = \preceq_{\overline{K}}^T$, so

$$\begin{aligned}\max(B_K, \sqsubseteq_K^T) &= \max(A_{\overline{K}}, \preceq_{\overline{K}}^T) \\ &= \operatorname*{arg\,max}_{b \in A_{\overline{K}}} |\overline{K}(b)| \\ &= \max(A_{\overline{K}}, \preceq_{\overline{K}}^{T_{\mathsf{CI}}}) \\ &= \max(B, \sqsubseteq_K^{T_{\mathsf{CI}}})\end{aligned}$$

where we apply **Argmax** for T and T_{CI} to the dual tournament $\overline{K}$. This completes the proof. $\square$

3.6 Related Work

Chain graphs. Chain graphs were originally introduced by Yannakakis (1981), who proved that *chain completion* – finding the minimum number of edges that when added to a bipartite graph form a chain graph – is NP-complete. Hardness results have subsequently been obtained for chain *deletion* (Natanzon, Shamir, and Sharan 2001) (where only edge deletions are allowed) and chain *editing* (Drange et al. 2015) (where both additions and deletions are allowed). We refer the reader to the work of Jiao, Ravi, and Gatterbauer (2017) and Drange et al. (2015) for a more detailed account of this literature. Outside of complexity theory, chain graphs have been studied for their spectral properties in (Anđelić et al. 2015; Ghorbani 2017), and the more general notion of a *nested colouring* was introduced in (Cook II 2015).

Tournaments in social choice. Tournaments have important applications in the design of voting rules, where an alternative x beats y in a pairwise comparison if a majority of voters prefer x to y. Various *tournament solutions* have been proposed, which select a set of "winners" from a given tournament.[19] Of particular relevance to our work are the *Slater set* and *Kemeny's rule* (Brandt, Brill, et al. 2016), which find minimal sets of edges to invert in the tournament graph such that the beating relation becomes a total order.[20] These methods are intuitively similar to chain editing: both involve making minimal changes to the tournament until some property is satisfied. A rough analogue to the Slater set in our framework is the union of the top-ranked players from each $K' \in \mathcal{M}(K)$. Solutions based on the covering relation – such as the *uncovered* and *Banks* set (Brandt, Brill, et al. 2016) – also bear similarity to chain editing.

Finally, note that directed versions of chain graphs (obtained by orienting edges from A to B and adding missing edges from B to A) correspond to *acyclic tournaments*, and a topological sort of A becomes a linearisation of the chain ranking $\leqslant_K^A$. This suggests a connection between chain deletion and the standard *feedback arc set* problem for removing cycles and obtaining a ranking.

Generalised tournaments. A *generalised* tournament (González-Díaz, Hendrickx, and Lohmann 2014) is a pair (X, A), where $X = [t]$ for some $t \in \mathbb{N}$ and $A \in \mathbb{R}_{\geq 0}^{t \times t}$ is a non-negative $t \times t$ matrix with $A_{ii} = 0$ for all $i \in X$. In this formalism each encounter between a pair of players i and j is represented by *two* numbers: A_{ij} and A_{ji}. This allows one to model both intensities of victories and losses (including draws) via the difference $A_{ij} - A_{ji}$, and the case where a comparison is not available (where $A_{ij} = A_{ji} = 0$).

Any $m \times n$ bipartite tournament K has a natural generalised tournament representation via the $(m+n) \times (m+n)$ *anti-diagonal block matrix* $A = \left[\begin{smallmatrix} 0 & K \\ \bar{K} & 0 \end{smallmatrix}\right]$, where the top-left and bottom-right blocks are the $m \times m$ and $n \times n$ zero matrices respectively. However, such anti-diagonal block matrices are often excluded in the generalised tournament literature due to an assumption of *irreducibility*, which requires that the directed graph corresponding to A is strongly connected. This is not the case in general for A constructed as above, which means not all existing tournament operators (and tournament axioms)

[19]Note that a ranking, such as we consider in this chapter, induces a set of winners by taking the maximally ranked players.

[20]Note that like chain editing, Kemeny's rule also admits a maximum likelihood characterisation (Elkind and Slinko 2016).

are well-defined for bipartite inputs.[21] Consequently, bipartite tournaments are a special case of generalised tournaments *in principle*, but not in practise.

3.7 Conclusion

Summary. In this chapter we studied chain editing, an interesting problem from computational complexity theory, as a ranking mechanism for bipartite tournaments. We analysed such mechanisms from a probabilistic viewpoint via the MLE characterisation, and in axiomatic terms. To resolve both the failure of an important anonymity axiom and NP-hardness, we weakened the chain editing requirement to one of *chain definability*, and characterised the resulting class of operators by the intuitive interleaving algorithm. Moreover, we characterised the particular interleaving instance T_{CI} by way of new axioms.

Limitations and future work. The hardness of chain editing remains a limitation of our approach. A possible remedy is to look to one of the numerous variant problems that are polynomial-time solvable. For instance, Jiao, Ravi, and Gatterbauer (2017) found a polynomial-time algorithm for *constrained k-near chain editing*, where the ranking of B is fixed, and one must minimise edge edits to find a chain graph in which each $a \in A$ deviates from a given fixed ranking by no more than k positions. This scenario could be relevant to our problem if prior rankings are available (for instance, from a previous tournament). One could develop approximation algorithms for chain editing, possibly based on existing approximations of chain completion (Natanzon, Shamir, and Sharan 2000). The interleaving operators of Section 3.5.2 go in this direction, but we did not yet obtain any theoretical or experimental bounds on the approximation ratio.

A second limitation of our work lies in the assumptions of the probabilistic model; namely that the true state of the world can be reduced to vectors of numerical skill levels which totally describe the tournament participants. This assumption may be violated when the competitive element of a tournament is *multi-faceted*, since a single number cannot represent multiple orthogonal components of a player's capabilities. For instance, in the truth discovery scenario of Example 3.0.1 sources may have varying levels of trustworthiness across different topics.[22] Nevertheless, if skill levels are taken as *aggregations* of these components, chain editing may prove to be a useful, albeit simplified, model.

The basic assumptions on tournaments could also be relaxed to bring our setting closer to generalised tournaments; e.g. by considering *partial tournaments*, in which not every a faces every b, and *graded outcomes*, in which

[21] We note that Slutzki and Volij (2005) side-step the reducibility issue by decomposing T into irreducible components and ranking each separately, although their methods may give only *partial* orders.

[22] Topic-dependent expertise will be considered in the next chapter.

results are not limited to 0 or 1 (indicating a win for a or b, respectively), but may take any value in the interval $[0, 1]$. Such extensions would improve the applicability of bipartite tournaments to new domains. The question of how best to adapt chain editing to these cases remains open.

Finally, there is room for more detailed axiomatic investigation. In this chapter we have stuck with fairly standard social choice axioms and performed preliminary analysis. However, the indirect nature of the comparisons in a bipartite tournament presents unique challenges; new axioms may need to be formulated to properly evaluate bipartite ranking methods in a normative sense.

Interlude

In Chapters 2 and 3 we studied how to rank sources by trustworthiness, broadly speaking, in the context of truth discovery and bipartite tournament ranking. However, we had no formal semantics to define the *meaning* of trustworthiness, and indeed this meaning varies between truth discovery operators and ranking methods. This flexibility was useful for our social-choice-style analysis, where rankings are commonly used in this manner. For instance, truth discovery operators using different interpretations of trustworthiness can still be meaningfully compared by the axiomatic method.

In the remainder of the thesis we take a stricter view on trustworthiness, positioned in relation to *expertise* in a logic-based framework. Informally, a source is trustworthy with respect to a topic X if we *believe they are an expert* on matters relating to X. Under this interpretation, the truth discovery problem can be modelled in a similar way to *belief revision* (Alchourrón, Gärdenfors, and Makinson 1985) and *belief merging* (Konieczny and Pino Pérez 2002) by considering how to form *beliefs* on the basis of input reports. Specifically, by considering beliefs both about the state of the world and the expertise of the sources, we have analogues of both the source and claim rankings from truth discovery. Taking such a view grounds trust in formal semantics, and allows us to take a deep dive into how this notion of trust interacts with belief.

Beyond its relation to trustworthiness, expertise is also a topic of interest in its own right. The logical properties of expertise – and the connections to the truthfulness of information – are explored in detail in the next chapter using the tools of modal logic. This framework is used as the basis for a belief change problem in Chapter 5, which can be thought of as a complementary, logical version of truth discovery. Finally, Chapter 6 shifts the focus away from normative properties of aggregation methods – as expressed by axioms in Chapter 2 and postulates in Chapter 5 – towards *truth-tracking*, i.e. how one can find the truth given reports from non-expert sources.

— 4 —

Expertise and Information

In order to properly assess incoming information, it is important to consider the expertise of the reporting source. We should generally believe statements within the domain of expertise of the source, but ignore (or otherwise discount) statements about which the source has no expertise. This applies even when dealing with honest sources: a well-meaning but non-expert source may make false claims due to lack of expertise on the relevant facts. The situation may be further complicated if a source comments on multiple topics at once: we must *filter out* the parts of the statement within their domain of expertise.

Problems associated with expertise have been exacerbated recently by the COVID-19 pandemic, in which false information from non-experts has been shared widely on social media (Llewellyn 2020; van Dijck and Alinejad 2020). There have also been high-profile instances of experts going beyond their area of expertise to comment on issues of public health (Xaudiera and Cardenal 2020), highlighting the importance of *domain-specific* notions of expertise. Identifying experts is also an important task for *liquid democracy* (Blum and Zuber 2016), in which voters may delegate their votes to expertise on a given policy issue.

Expertise has been well-studied, with perspectives from behavioural and cognitive science (Chi, Glaser, and Farr 2014; Ericsson and Towne 2010), sociology (Collins and Evans 2008), and philosophy (Kilov 2021; Whyte and Crease 2010; Goldman 2018), among other fields. In this work we study the *logical* content of expertise, and its relation to truthfulness of information.

Specifically, we develop a *modal logic* framework to model *expertise* and *soundness of information*. Intuitively, a source has expertise on φ if they are able to correctly refute φ in any situation where it is false.[1] Thus, our notion of expertise *does not depend on the "actual" state of affairs*, but only on the source's epistemic state.

It is *sound* for a source to report φ if φ is true *up to lack of expertise*: if φ is logically weakened to a proposition ψ on which the source has expertise, then ψ must be true. That is, the consequences of φ on which the source has expertise are true. This formalises the idea of "filtering out" parts of a

[1] Note that we could instead consider the dual case: expertise means being able to *verify* when a proposition is true.

statement within a source's expertise. For example, suppose $\varphi = p \land q$, and the source has expertise on p but not q. Supposing p is true but q is false, φ is false. However, if we discard information by ignoring q (on which the source has no expertise), we obtain the weaker formula p, on which the source *does* have expertise, and which is true. If this holds for all possible ways to weaken $p \land q$ (this is the case, for instance, if the source does not have expertise on any statement strictly stronger than p), then $p \land q$ is *false* but *sound* for the source to report. In terms of refutation, φ is sound if the source cannot refute $\neg\varphi$. That is, either φ is in fact true, or the source does not possess sufficient expertise to rule out φ.

This informal picture of expertise already suggests a close connection between expertise, soundness and *knowledge*. Indeed, we will see that, under certain conditions, expertise can be equivalently interpreted in terms of *S4 or S5 knowledge*, familiar from epistemic logic.

Beyond the individual expertise of a single source, one can also consider the *collective expertise* of a group. For example, a committee may consist of several experts across different domains, so that by working together the group achieves expertise beyond any of its individual members. Indeed, such pooling of expertise becomes necessary in cases where it is infeasible for an individual to be a specialist in all relevant sub-areas. As a concrete example, consider the *Rogers Commission report*[2] into the 1986 Challenger disaster, whose members included politicians, military generals, physicists, astronauts and rocket scientists. Beyond extending the expertise of its constituents, the breadth of expertise among the commission allowed it to collectively assess issues at the *intersection* of its members' specialities.

Towards defining collective expertise we will again turn to (multi-agent) epistemic logic, borrowing from the well-known notions of *distributed* and *common knowledge* (Fagin et al. 2003). Just as individual expertise (and soundness) can be expressed in terms of knowledge, we will see that collective expertise can be expressed in terms of collective knowledge.

While the picture of expertise painted so far has been static, it is also natural to consider the *dynamics* of expertise. For example, how does expertise change over time as sources interact with the world and gain new knowledge? What are the effects of *announcements*, particularly when sources are non-experts? We study the logic of such events via dynamic operators in the style of dynamic epistemic logic (van Ditmarsch, van der Hoek, and Kooi 2008), and particularly *dynamic evidence logics* (van Benthem and Pacuit 2011; van Benthem, Fernández-Duque, and Pacuit 2014).

Contributions. On the conceptual side, we develop a modal logic framework to reason about the expertise of a source and soundness of information. We also study collective expertise among multiple sources, and consider how expertise may evolve via learning and announcements. Importantly, both singular and

[2] https://en.wikipedia.org/wiki/Rogers_Commission_Report

collective expertise are shown to be connected in a precise sense to standard notions from epistemic logic. This formalises the conceptual link between expertise and *knowledge*. On the technical side we obtain a sound and complete axiomatisation, and axiomatise several sub-classes of models with additional axioms.

This chapter is an extension of Singleton and Booth (2023) and Singleton (2021), with new material appearing in Section 4.6. Several of the main proofs have been formalised with the Lean theorem prover.[3]

4.1 Expertise and Soundness

Before the formal definitions we give an example to illustrate the notions of *expertise* and *soundness*, which are central to the framework.

Example 4.1.1. *Consider an economist reporting on the possible impact of a novel virus which has recently been detected. The virus may or may not be highly infectious (i) and go on to cause a high death toll (d), and there may or may not be economic prosperity in the near future (p). The economist reports that despite the virus, the economy will prosper and there will not be mass deaths ($p \land \neg d$). Assume the economist is an expert on matters relating to the economy ($\mathsf{E}p$, $\mathsf{E}\neg p$), but not on matters of public health ($\neg\mathsf{E}d$, $\neg\mathsf{E}\neg d$). For the sake of the example, suppose the virus will in fact cause a high death toll, but the economy will nonetheless prosper. Then while the report of $p \land \neg d$ is false, it is true if one* ignores the parts on which the economist has no expertise *(namely, $\neg d$); in doing so we obtain p, which is true. The report therefore carries* some *true information, even though it is false. We say $p \land \neg d$ is* sound *for the economist in this case.*

Syntax. Let Prop be a countable set of atomic propositions. To start with, we consider a single information source. Our language $\mathcal{L}$ includes modal operators to express expertise and soundness statements for this source, and is defined by the following grammar:

$$\varphi ::= p \mid \varphi \land \varphi \mid \neg\varphi \mid \mathsf{E}\varphi \mid \mathsf{S}\varphi \mid \mathsf{A}\varphi$$

for $p \in \mathsf{Prop}$. We read $\mathsf{E}\varphi$ as "the source has expertise on φ, and $\mathsf{S}\varphi$ as "φ is sound for the source to report". We include the universal modality A for technical convenience; $\mathsf{A}\varphi$ is read as "φ holds in all states" (Goranko and Passy 1992). Other logical connectives ($\lor$, $\to$, $\leftrightarrow$) and constants ($\top$, $\bot$) are introduced as abbreviations.

Semantics. On the semantic side, we use the notion of an *expertise model.*

[3] https://github.com/joesingo/expertise-and-information

Definition 4.1.1. *An* expertise model *(hereafter, just* model*) is a triple $M = (X, P, V)$, where X is a set of states, $P \subseteq 2^X$ is a collection of subsets of X, and $V : \mathsf{Prop} \to 2^X$ is a valuation function. An* expertise frame *is a pair $F = (X, P)$. The class of all models is denoted by $\mathbb{M}$.*

The sets in P are termed *expertise sets*, and represent the propositions on which the source has expertise. Given the earlier informal description of expertise as refutation, we interpret $A \in P$ as saying that whenever the "actual" state is outside A, the source knows so.

For an expertise model $M = (X, P, V)$, the satisfaction relation between states $x \in X$ and formulas $\varphi \in \mathcal{L}$ is defined recursively as follows:

$$\begin{array}{lll} M, x & \models p & \iff x \in V(p) \\ M, x & \models \varphi \wedge \psi & \iff M, x \models \varphi \text{ and } M, x \models \psi \\ M, x & \models \neg\varphi & \iff M, x \not\models \varphi \\ M, x & \models \mathsf{E}\varphi & \iff \|\varphi\|_M \in P \\ M, x & \models \mathsf{S}\varphi & \iff \forall A \in P : \|\varphi\|_M \subseteq A \implies x \in A \\ M, x & \models \mathsf{A}\varphi & \iff \forall y \in X : M, y \models \varphi \end{array}$$

where $\|\varphi\|_M = \{x \in X \mid M, x \models \varphi\}$ is the truth set of φ. For an expertise frame $F = (X, P)$, write $F \models \varphi$ iff $M, x \models \varphi$ for all models M based on F and all $x \in X$. Write $M \models \varphi$ iff $M, x \models \varphi$ for all $x \in X$, and $\models \varphi$ iff $M \models \varphi$ for all models M; we say φ is *valid* in this case. Write $\varphi \equiv \psi$ iff $\varphi \leftrightarrow \psi$ is valid. For a set $\Gamma \subseteq \mathcal{L}$, write $\Gamma \models \varphi$ iff for all models M and states x, if $M, x \models \psi$ for all $\psi \in \Gamma$ then $M, x \models \varphi$.

The clauses for atomic propositions and propositional connectives are standard. For expertise formulas, we have that $\mathsf{E}\varphi$ holds exactly when the set of states where φ is true is an element of P. Expertise is thus a special case of the *neighbourhood semantics* (Scott 1970; Montague 1970; Pacuit 2017), where each point $x \in X$ has the same set of neighbourhoods. The clause for soundness reflects the intuition that φ is sound exactly when all logically weaker formulas on which the source has expertise must be true: if $A \in P$ (i.e. the source has expertise on A) and A contains all φ states, then $x \in A$. In terms of refutation, $\mathsf{S}\varphi$ holds iff there is no expertise set A, false at the actual state x, which allows the source to rule out φ.

Our truth conditions for expertise and soundness also have topological interpretations, if one views P as the collection of closed sets of a topology on X:[4] $\mathsf{E}\varphi$ holds iff $\|\varphi\|_M$ is closed, and $\mathsf{S}\varphi$ holds at x iff x lies in the *closure* of $\|\varphi\|_M$. Our semantics for soundness is therefore dual to the *interior semantics* for modal logic, where $\Box\varphi$ is true at x iff x lies in the interior of $\|\varphi\|$. We can also view the closure operation as *expanding* the set $\|\varphi\|_M$ along the lines of the source's expertise; φ is sound if the "actual" state x is included in this

[4]For this to be the case, P must be closed under intersections and finite unions, and contain both the empty set and X itself. We will turn to these closure properties in Section 4.2.

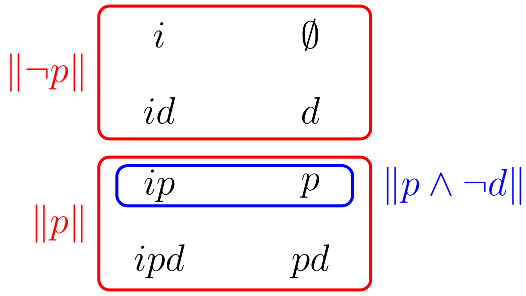


Figure 4.1: Expertise model from Example 4.1.2, which formalises Example 4.1.1.

expansion. Finally, the clause for the universal modality A states that $\mathsf{A}\varphi$ holds iff φ holds at all states $y \in X$.

Example 4.1.2. *To formalise Example 4.1.1, consider the model $M = (X, P, V)$ shown in Fig. 4.1, where $X = 2^{\{i,p,d\}}$, $P = \{\{ipd, pd, ip, p\}, \{id, d, i, \emptyset\}\}$ (indicated by the solid rectangles; sets in X are written as strings for brevity), and $V(q) = \{S \mid q \in S\}$. Then we have $M \models \mathsf{E}p$ but $M \not\models \mathsf{E}d$. The economist's report of $p \wedge \neg d$ is represented by the blue region. We see that while $M, ipd \not\models p \wedge \neg d$, all expertise sets containing the blue region also contain ipd, so $M, ipd \models \mathsf{S}(p \wedge \neg d)$. That is, the economist's report is false but sound if the "actual" state of the world were ipd. This act of "expanding" $\|p \wedge \neg d\|$ until we reach an expertise set corresponds to ignoring the parts of the report on which the economist has no expertise, as in Example 4.1.1.*

We further illustrate the semantics by listing some valid formulas.

Proposition 4.1.1. *The following formulas are valid:*

1. $\varphi \rightarrow \mathsf{S}\varphi$
2. $\mathsf{E}\varphi \leftrightarrow \mathsf{AE}\varphi$
3. $\mathsf{A}(\varphi \rightarrow \psi) \rightarrow (\mathsf{S}\varphi \wedge \mathsf{E}\psi \rightarrow \psi)$
4. $\mathsf{E}\varphi \rightarrow \mathsf{A}(\mathsf{S}\varphi \rightarrow \varphi)$

Proof. Let $M = (X, P, V)$ be a model and $x \in X$. (1) and (2) are clear. For (3), suppose $M, x \models \mathsf{A}(\varphi \rightarrow \psi)$. Then $\|\varphi\|_M \subseteq \|\psi\|_M$. Further, suppose $M, x \models \mathsf{S}\varphi \wedge \mathsf{E}\psi$. Then $\|\varphi\|_M \subseteq \|\psi\|_M \in P$; taking $A = \|\psi\|_M$ in the definition of the semantics for S, we get by $M, x \models \mathsf{S}\varphi$ that $x \in \|\psi\|_M$, i.e. $M, x \models \psi$. Finally, (4) follows from (2) and (3) by taking $\psi = \varphi$. □

Here (1) says that all truths are sound. (2) says that expertise is global. (3) says that if the source has expertise on ψ, and ψ is logically weaker than some sound formula φ, then ψ is in fact true. This formalises the idea that if φ is true *up to lack of expertise*, then weakening φ until expertise holds (i.e.

[4]

discarding parts of φ on which the source does not have expertise) results in something true. (4) says that if the source has expertise on φ, then whenever φ is sound it is also true.

4.2 Closure Properties

So far we have not imposed any constraints on the collection of expertise sets P. But given our interpretation of P, it may be natural to require that P is closed under certain set-theoretic operations. Say a frame $F = (X, P)$ is

- *closed under intersections* if $\{A_i\}_{i \in I} \subseteq P$ implies $\bigcap_{i \in I} A_i \in P$
- *closed under unions* if $\{A_i\}_{i \in I} \subseteq P$ implies $\bigcup_{i \in I} A_i \in P$
- *closed under finite unions* if $A, B \in P$ implies $A \cup B \in P$
- *closed under complements* if $A \in P$ implies $X \setminus A \in P$

In the first two cases we allow the empty collection $\emptyset \subseteq P$, and employ the nullary intersection convention $\bigcap \emptyset = X$. Consequently, closure under intersections implies $X \in P$, and closure under unions implies $\emptyset \in P$.

Say a model has any of the above properties if the underlying frame does. Write $\mathbb{M}_{\mathsf{int}}$, $\mathbb{M}_{\mathsf{unions}}$, $\mathbb{M}_{\mathsf{unions}}$, $\mathbb{M}_{\mathsf{finite-unions}}$ and $\mathbb{M}_{\mathsf{compl}}$ for the classes of models closed under intersections, unions, finite unions and complements respectively.

What are the intuitive interpretations of these closure conditions? Consider again our interpretation of $A \in P$: whenever the actual state is not in A, the source knows so. With this in mind, closure under intersections is a natural property: if $x \notin \bigcap_{i \in I} A_i$ then there is some $i \in I$ such that $x \notin A_i$; the source can then use this to refute A_i and therefore know that the actual state x does not lie in the intersection $\bigcap_{i \in I} A_i$. A similar argument can be made for finite unions: if $x \notin A \cup B$ then the source can use $x \notin A$ and $x \notin B$ to refute both A and B. Closure under *arbitrary* unions is less clear cut; determining that $x \notin \bigcup_{i \in I} A_i$ requires the source to refute (potentially) infinitely many propositions A_i. This is more demanding from a computational and cognitive perspective, and we therefore view closure under (arbitrary) unions as an optional property which may or may not be appropriate depending on the situation one wishes to model. Finally, closure under complements removes the distinction between refutation and *verification*: if the agent can refute A whenever A is false, they can also verify A whenever A is true. We view this as another optional property, which is appropriate in situations where *symmetric* expertise is desirable (i.e. when expertise on φ and $\neg\varphi$ should be considered equivalent).

Several of these properties can be formally captured in our language at the level of frames.

Proposition 4.2.1. *Let $F = (X, P)$ be a non-empty frame. Then*

1. *F is closed under intersections iff $F \models \mathsf{A}(\mathsf{S}\varphi \to \varphi) \to \mathsf{E}\varphi$ for all $\varphi \in \mathcal{L}$*

2. *F is closed under finite unions iff $F \models \mathsf{E}\varphi \wedge \mathsf{E}\psi \to \mathsf{E}(\varphi \vee \psi)$ for all $\varphi, \psi \in \mathcal{L}$*

3. *F is closed under complements iff $F \models \mathsf{E}\varphi \leftrightarrow \mathsf{E}\neg\varphi$ for all $\varphi \in \mathcal{L}$*

Proof. We prove only the first claim; the others are straightforward.

"if": We show the contrapositive. Suppose F is not closed under intersections. Then there is a collection $\{A_i\}_{i\in I} \subseteq P$ such that $B := \bigcap_{i\in I} A_i \notin P$. Let p be an arbitrary atomic proposition, and define a valuation V by $V(p) = B$ and $V(q) = \emptyset$ for $q \neq p$. Let $M = (X, P, V)$ be the corresponding model. Since X is assumed to be non-empty, we may take some $x \in X$.

We claim that $M, x \models \mathsf{A}(\mathsf{S}p \to p)$ but $M, x \not\models \mathsf{E}p$. Clearly $M, x \not\models \mathsf{E}p$ since $\|p\|_M = B \notin P$. For $M, x \models \mathsf{A}(\mathsf{S}p \to p)$, suppose $y \in X$ and $M, y \models \mathsf{S}p$. Let $j \in I$. Then $A_j \in P$, and

$$\|p\|_M = B = \bigcap_{i\in I} A_i \subseteq A_j$$

so by $M, y \models \mathsf{S}p$ we have $y \in A_j$. Hence $y \in \bigcap_{j\in I} A_j = B = \|p\|_M$, so $M, y \models p$. This shows that any $y \in X$ has $M, y \models \mathsf{S}p \to p$, and thus $M, x \models \mathsf{A}(\mathsf{S}p \to p)$. Hence $F \not\models \mathsf{A}(\mathsf{S}p \to p) \to \mathsf{E}p$.

"only if": Suppose F is closed under intersections. Let M be a model based on F and take $x \in X$. Let $\varphi \in \mathcal{L}$. Suppose $M, x \models \mathsf{A}(\mathsf{S}\varphi \to \varphi)$. Then $\|\mathsf{S}\varphi\|_M \subseteq \|\varphi\|_M$. But since $\models \varphi \to \mathsf{S}\varphi$, we have $\|\varphi\|_M \subseteq \|\mathsf{S}\varphi\|_M$ too. Hence $\|\varphi\|_M = \|\mathsf{S}\varphi\|_M$, i.e.

$$\|\varphi\|_M = \|\mathsf{S}\varphi\|_M = \bigcap\{A \in P \mid \|\varphi\|_M \subseteq A\} \in P$$

where we use the fact that P is closed under intersections in the final step. Hence $\|\varphi\|_M \in P$, so $M, x \models \mathsf{E}\varphi$. □

The question of whether closure under (arbitrary) unions can be expressed in the language is still open. By Proposition 4.2.1 (1) and Proposition 4.1.1 (4), the language fragment $\mathcal{L}_{\mathsf{SA}}$ containing only the S and A modalities is equally expressive as the full language $\mathcal{L}$ with respect to $\mathbb{M}_{\mathsf{int}}$, since $\mathsf{E}\varphi$ is equivalent to $\mathsf{A}(\mathsf{S}\varphi \to \varphi)$ in such models. In general $\mathcal{L}_{\mathsf{SA}}$ is strictly less expressive, since $\mathcal{L}_{\mathsf{SA}}$ cannot distinguish between a model and its closure under intersections.

Lemma 4.2.1. *Let $M = (X, P, V)$ be a model, and $M' = (X, P', V)$ its closure under intersections, where $A \in P'$ iff $A = \bigcap_{i\in I} A_i$ for some $\{A_i\}_{i\in I} \subseteq P$. Then for all $\varphi \in \mathcal{L}_{\mathsf{SA}}$ and $x \in X$, we have $M, x \models \varphi$ iff $M', x \models \varphi$.*

Proof. By induction on $\mathcal{L}_{\mathsf{SA}}$ formulas. The cases for atomic propositions, propositional connectives and A are straightforward. We treat only the case for S. The "if" direction is clear using the induction hypothesis and the fact that $P \subseteq P'$. Suppose $M, x \models \mathsf{S}\varphi$. Take $A = \bigcap_{i \in I} A_i \in P'$, where each A_i is in P, such that $\|\varphi\|_{M'} \subseteq A$. By the induction hypothesis, $\|\varphi\|_M \subseteq A$. For any $i \in I$, $\|\varphi\|_M \subseteq A \subseteq A_i$ and $M, x \models \mathsf{S}\varphi$ gives $x \in A_i$. Hence $x \in \bigcap_{i \in I} A_i = A$. This shows $M', x \models \mathsf{S}\varphi$. $\square$

It follows that $\mathcal{L}_{\mathsf{SA}}$ is strictly less expressive than $\mathcal{L}$.[5] To round off the discussion of closure properties, we note that within the class of frames closed under intersections, closure under finite unions is also captured by the well-known **K** axiom – $\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$ – for the dual soundness operator $\hat{\mathsf{S}}\varphi := \neg\mathsf{S}\neg\varphi$:

Proposition 4.2.2. *Suppose $F = (X, P)$ is non-empty and closed under intersections. Then F is closed under finite unions if and only if $F \models \hat{\mathsf{S}}(\varphi \to \psi) \to (\hat{\mathsf{S}}\varphi \to \hat{\mathsf{S}}\psi)$ for all $\varphi, \psi \in \mathcal{L}$.*

Proof. "if": We show the contrapositive. Suppose F is closed under intersections but not finite unions, so that there are $B_1, B_2 \in P$ with $B_1 \cup B_2 \notin P$. Set

$$C = \bigcap \{A \in P \mid B_1 \cup B_2 \subseteq A\}$$

By closure under intersections, $C \in P$. Clearly $B_1 \cup B_2 \subseteq C$. Since $C \in P$ but $B_1 \cup B_2 \notin P$, $B_1 \cup B_2 \subset C$. Hence there is $x \in C \setminus (B_1 \cup B_2)$.

Now pick distinct atomic propositions p and q, and let V be any valuation with $V(p) = B_1 \cup B_2$ and $V(q) = B_1$. Let $M = (X, P, V)$ be the corresponding model. We make three claims:

- $M, x \models \mathsf{S}p$: Take $A \in P$ such that $\|p\|_M \subseteq A$. Then $B_1 \cup B_2 \subseteq A$, so $C \subseteq A$. Since $x \in C$, we have $x \in A$ as required.
- $M, x \not\models \mathsf{S}q$: This is clear since $B_1 \in P$, $\|q\|_M \subseteq B_1$, but $x \notin B_1$.
- $M, x \not\models \mathsf{S}(p \wedge \neg q)$: Note that $\|p \wedge \neg q\|_M = V(p) \setminus V(q) = B_2 \setminus B_1$. Therefore we have $B_2 \in P$ and $\|p \wedge \neg q\|_M \subseteq B_2$, but $x \notin B_2$.

Now set $\varphi = \neg q$ and $\psi = \neg p$. We have

$$\hat{\mathsf{S}}(\varphi \to \psi) = \neg\mathsf{S}\neg(\varphi \to \psi) \equiv \neg\mathsf{S}(\varphi \wedge \neg\psi) \equiv \neg\mathsf{S}(p \wedge \neg q)$$

$$\hat{\mathsf{S}}\varphi \to \hat{\mathsf{S}}\psi = \neg\mathsf{S}\neg\varphi \to \neg\mathsf{S}\neg\psi \equiv \neg\mathsf{S}q \to \neg\mathsf{S}p \equiv \mathsf{S}p \to \mathsf{S}q$$

From the claims above we see that $M, x \models \hat{\mathsf{S}}(\varphi \to \psi)$ but $M, x \not\models \hat{\mathsf{S}}\varphi \to \hat{\mathsf{S}}\psi$. Since M is a model based on F, we are done.

"only if": Suppose F is closed under intersections and finite unions. Let M be a model based on F and x a state in M. Suppose $M, x \models \hat{\mathsf{S}}(\varphi \to \psi)$ and

[5] Indeed, consider $M = (X, P, V)$, where $X = \{1, 2, 3\}$, $P = \{\{1, 2\}, \{2, 3\}\}$ and $V(p) = \{1, 2\}$, $V(q) = \{2, 3\}$ for some fixed $p, q \in \mathsf{Prop}$. Let M' be as in Lemma 4.2.1. Then $M', 1 \models \mathsf{E}(p \wedge q)$ and $M, 1 \not\models \mathsf{E}(p \wedge q)$, but M and M' agree on $\mathcal{L}_{\mathsf{SA}}$ formulas. Hence $\mathsf{E}(p \wedge q)$ is not equivalent to any $\mathcal{L}_{\mathsf{SA}}$ formula.

$M, x \models \hat{\mathsf{S}}\varphi$. Then $M, x \not\models \mathsf{S}\neg(\varphi \to \psi)$ and $M, x \not\models \mathsf{S}\neg\varphi$. Hence there is $A \in P$ such that $\|\neg(\varphi \to \psi)\|_M \subseteq A$ but $x \notin A$, and $B \in P$ such that $\|\neg\varphi\|_M \subseteq B$ but $x \notin B$. Note

$$\|\neg\psi\|_M \subseteq \|\varphi \wedge \neg\psi\|_M \cup \|\neg\varphi\|_M = \|\neg(\varphi \to \psi)\|_M \cup \|\neg\varphi\|_M \subseteq A \cup B.$$

Since $x \notin A \cup B$ and $A \cup B \in P$ by closure under finite unions, this shows $M, x \not\models \mathsf{S}\neg\psi$, i.e. $M, x \models \hat{\mathsf{S}}\psi$. This completes the proof of $F \models \hat{\mathsf{S}}(\varphi \to \psi) \to (\hat{\mathsf{S}}\varphi \to \hat{\mathsf{S}}\psi)$. □

4.3 Connection with Epistemic Logic

In this section we explore the connection between our logic and *epistemic logic*, for certain classes of expertise models. In particular, we show a one-to-one mapping between classes of expertise models and *S4 and S5 relational models*, and a translation from $\mathcal{L}$ to the modal language with knowledge operator K which allows expertise and soundness to be expressed in terms of *knowledge*.

First, we introduce the syntax and (relational) semantics of epistemic logic. Let $\mathcal{L}_{\mathsf{KA}}$ be the language formed from Prop with modal operators K and A. We read $\mathsf{K}\varphi$ as *the source knows* φ.

Definition 4.3.1. *A* relational model *is a triple $M^* = (X, R, V)$, where X is a set of states, $R \subseteq X \times X$ is a binary relation on X, and $V : \mathsf{Prop} \to 2^X$ is a valuation function. The class of all relational models is denoted by $\mathbb{M}^*$.*

The satisfaction relation for $\mathcal{L}_{\mathsf{KA}}$ is defined recursively: the clauses for atomic propositions, propositional connectives and A are the same as for expertise models, and

$$M^*, x \models \mathsf{K}\varphi \iff \forall y \in X : xRy \implies M^*, y \models \varphi.$$

As is standard, R is interpreted as an *epistemic accessibility relation*: xRy means that the source considers y possible if the "actual" state of the world is x. We will be interested in the logics of S4 and S5, which are axiomatised by **KT4** and **KT5**, respectively:

- **K**: $\mathsf{K}(\varphi \to \psi) \to (\mathsf{K}\varphi \to \mathsf{K}\psi)$
- **T**: $\mathsf{K}\varphi \to \varphi$
- **4**: $\mathsf{K}\varphi \to \mathsf{K}\mathsf{K}\varphi$
- **5**: $\neg\mathsf{K}\varphi \to \mathsf{K}\neg\mathsf{K}\varphi$

T says that all knowledge is true, **4** expresses *positive introspection* of knowledge, and **5** expresses *negative introspection*. One can show that **K**, **T** and **5** together prove **4** (Zach 2019, p. 51), and some authors write **KT45**

instead of **KT5**. S5 is therefore *stronger* than S4, in the sense that any formula provable in S4 is also provable in S5. In fact, S5 has been criticised as *too* strong in the philosophical literature, since the negative introspection **5** is a rather idealised property of knowledge. For example, it is certainly reasonable to expect that **5** may fail for agents who are not perfectly rational (e.g. humans).

It is well known that S4 is sound and complete for the class of relational models where R is reflexive and transitive, and that S5 is sound and complete for the class of relational models where R is an equivalence relation. Accordingly, we write $\mathbb{M}^*_{\mathsf{S4}}$ for the class of all M^* where R is reflexive and transitive, and $\mathbb{M}^*_{\mathsf{S5}}$ for M^* where R is an equivalence relation.

Our first result connecting expertise and knowledge is on the semantic side: we show there is a bijection between expertise models closed under intersections and unions and S4 models. Moreover, there is a close connection between the collection of expertise sets P and the corresponding relation R. Since expertise models closed under intersections and unions are *Alexandrov topological spaces* (where P is the set of closed sets), this is essentially a reformulation of a known result linking relational semantics over S4 frames and topological interior semantics over Alexandrov spaces (van Benthem and Bezhanishvili 2007; Özgün 2017).[6] To be self-contained, we prove it for our setting here. First, we show how to map a collection of sets P to a binary relation.

Definition 4.3.2. *For a set X and $P \subseteq 2^X$, let R_P be the binary relation on X given by*

$$xR_Py \iff \forall A \in P : (y \in A \implies x \in A).$$

In the case where P is the collection of closed sets of a topology on X, R_P is the *specialisation preorder*. Fig. 4.2 shows an example of R_P for X and P from Example 4.1.2. In what follows, say a set $A \subseteq X$ is *downwards closed* with respect to a relation R if xRy and $y \in A$ implies $x \in A$.

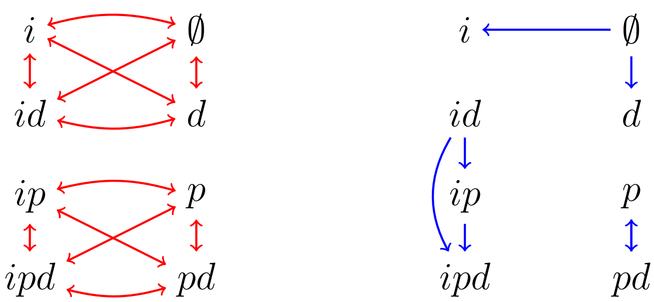


Figure 4.2: Left: the relation R_P corresponding to X and P from Example 4.1.2 (with reflexive edges omitted). Note that R_P is an equivalence relation, with equivalence classes $\|p\|$ and $\|\neg p\|$. Right: an example of a non-symmetric relation R_P, corresponding to $P = \{\emptyset, X, \{id, ip, ipd\}, \{id, ip\}, \{id\}, \{i, \emptyset\}, \{\emptyset, d\}, \{p, pd\}\}$.

[6] In fact, the interior semantics has an intrinsic epistemic interpretation (without appeal to any link with relational semantics) if one views open sets as *evidence* (Özgün 2017, pp. 24).

Lemma 4.3.1. *Let X be a set and R, S reflexive and transitive relations on X. Then if R and S share the same downwards closed sets, $R = S$.*

Proof. Suppose xRy. Set $A = \{z \in X \mid zSy\}$. By transitivity of S, A is downwards closed wrt S. By assumption, A must also be downwards closed wrt R. By reflexivity of S, $y \in A$. Hence xRy implies $x \in A$, i.e. xSy. This shows $R \subseteq S$, and the reverse inclusion holds by a symmetrical argument. Hence $R = S$. □

Lemma 4.3.2. *Let X be a set.*

1. *For any $P \subseteq 2^X$, R_P is reflexive and transitive.*
2. *If $P \subseteq 2^X$ is closed under unions and intersections, then for all $A \subseteq X$:*

$$A \in P \iff A \text{ is downwards closed wrt } R_P.$$

3. *If R is a reflexive and transitive relation on X, there is $P \subseteq 2^X$ closed under unions and intersections such that $R_P = R$.*

Proof.

1. Straightforward by the definition of R_P.
2. Suppose P is closed under unions and intersections and let $A \subseteq X$. First suppose $A \in P$. Then A is downwards closed with respect to R_P: if $y \in A$ and xR_Py then, by definition of R_P, we have $x \in A$.

   Next suppose A is downwards closed with respect to R_P. We claim

$$A = \bigcup_{y \in A} \bigcap \{B \in P \mid y \in B\}$$

   Since P is closed under intersections and unions, this will show $A \in P$. The left-to-right inclusion is clear, since any $y \in A$ lies in the term of the union corresponding to y. For the right-to-left inclusion, take any x in the set on the RHS. Then there is $y \in A$ such that $x \in \bigcap\{B \in P \mid y \in B\}$. But this is just a rephrasing of xR_Py. Since A is downwards closed, we get $x \in A$ as required.
3. Take any reflexive and transitive relation R. Set

$$P = \{A \subseteq X \mid A \text{ is downwards closed wrt } R\}.$$

   It is easily seen that P is closed under unions and intersections. We need to show that $R_P = R$. By (1), R_P is reflexive and transitive. By Lemma 4.3.1, it is sufficient to show that R_P and R share the same downwards closed sets. Indeed, for any $A \subseteq X$ we get by (2) and the definition of P that

$$\begin{aligned} A \text{ is downwards closed wrt } R_P &\iff A \in P \\ &\iff A \text{ is downwards closed wrt } R. \end{aligned}$$

Hence $R = R_P$.

$\square$

We can now state the correspondence between expertise models and S4 relational models.

Theorem 4.3.1. *The mapping* $f : \mathbb{M}_{\mathsf{int}} \cap \mathbb{M}_{\mathsf{unions}} \to \mathbb{M}^*_{\mathsf{S4}}$ *given by* $(X, P, V) \mapsto (X, R_P, V)$ *is bijective.*

Proof. Lemma 4.3.2 (1) shows that f is well-defined, i.e. that $f(M)$ does indeed lie in $\mathbb{M}^*_{\mathsf{S4}}$ for any expertise model M. Injectivity follows from Lemma 4.3.2 (2), since P is fully determined by R_P for expertise models closed under unions and intersections. Finally, Lemma 4.3.2 (3) shows that f is surjective. $\square$

If we consider closure under complements together with intersections, an analogous result holds with S5 taking the place of S4.

Theorem 4.3.2. *The mapping* $g : \mathbb{M}_{\mathsf{int}} \cap \mathbb{M}_{\mathsf{compl}} \to \mathbb{M}^*_{\mathsf{S5}}$ *given by* $(X, P, V) \mapsto (X, R_P, V)$ *is bijective.*

Proof. First, note that $\mathbb{M}_{\mathsf{int}} \cap \mathbb{M}_{\mathsf{compl}} \subseteq \mathbb{M}_{\mathsf{int}} \cap \mathbb{M}_{\mathsf{unions}}$, since any union of sets in P can be written as a complement of intersection of complements of sets in P. Therefore g is simply the restriction of f from Theorem 4.3.1 to $\mathbb{M}_{\mathsf{int}} \cap \mathbb{M}_{\mathsf{compl}}$.

To show g is well-defined, we need to show that R_P is an equivalence relation whenever P is closed under intersections and complements. Reflexivity and transitivity were already shown in Lemma 4.3.2 (1). We show R_P is symmetric. Suppose xR_Py. Let $A \in P$ such that $x \in A$. Write $B = X \setminus A$. Then since P is closed under complements, $B \in P$. Since xR_Py and $x \notin B$, we cannot have $y \in B$. Thus $y \notin B = X \setminus A$, i.e. $y \in A$. This shows yR_Px. Hence R_P is an equivalence relation.

Injectivity of g is inherited from injectivity of f from Theorem 4.3.1. For surjectivity, it suffices to show that $f^{-1}(M^*)$ is closed under complements when $M^* = (X, R, V) \in \mathbb{M}^*_{\mathsf{S5}}$. Recall, from Lemma 4.3.2 (3), that $f^{-1}(M^*) = (X, P, V)$, where $A \in P$ iff A is downwards closed with respect to R. Suppose $A \in P$, i.e. A is downwards closed. To show $X \setminus A$ is downwards closed, suppose $y \in X \setminus A$ and xRy. By symmetry of R, yRx. If $x \in A$, then downwards closure of A would give $y \in A$, but this is false. Hence $x \notin A$, i.e. $x \in X \setminus A$. Thus $X \setminus A$ is downwards closed, so P is closed under complements. This completes the proof. $\square$

The mappings between expertise models and relational models also preserve the truth value of formulas, via the following translation $t : \mathcal{L} \to \mathcal{L}_{\mathsf{KA}}$, which expresses expertise and soundness in terms of knowledge:

$$\begin{aligned}
t(p) &= p \\
t(\varphi \wedge \psi) &= t(\varphi) \wedge t(\psi) \\
t(\neg\varphi) &= \neg t(\varphi) \\
t(\mathsf{E}\varphi) &= \mathsf{A}(\neg t(\varphi) \to \mathsf{K}\neg t(\varphi)) \\
t(\mathsf{S}\varphi) &= \neg\mathsf{K}\neg t(\varphi) \\
t(\mathsf{A}\varphi) &= \mathsf{A}t(\varphi).
\end{aligned}$$

The only interesting cases are for $\mathsf{E}\varphi$ and $\mathsf{S}\varphi$. The translation of $\mathsf{E}\varphi$ corresponds directly to the intuition of expertise as refutation: in all possible scenarios, if φ is false the source knows so. The translation of $\mathsf{S}\varphi$ says that soundness is just the dual of knowledge: φ is sound if the source does not *know* that φ is false. Of particular interest is the case where φ lies in the purely propositional language $\mathcal{L}_0$, i.e. does not contain any modalities. Such formulas describe the "ontic" facts of the world, and do not refer to the expertise of the source. Since t leaves atomic propositions unchanged and preserves the structure of conjunctions and negations, we have $t(\varphi) = \varphi$. Consequently $t(\mathsf{E}\varphi) = \mathsf{A}(\neg\varphi \to \mathsf{K}\neg\varphi)$ and $t(\mathsf{S}\varphi) = \neg\mathsf{K}\neg\varphi$.

Theorem 4.3.3. *Let $f : \mathbb{M}_{\mathsf{int}} \cap \mathbb{M}_{\mathsf{unions}} \to \mathbb{M}^*_{\mathsf{S4}}$ be the bijection from Theorem 4.3.1. Then for all $M = (X, P, V) \in \mathbb{M}_{\mathsf{int}} \cap \mathbb{M}_{\mathsf{unions}}$, $x \in X$ and $\varphi \in \mathcal{L}$:*

$$M, x \models \varphi \iff f(M), x \models t(\varphi) \tag{4.1}$$

Moreover, if $g : \mathbb{M}_{\mathsf{int}} \cap \mathbb{M}_{\mathsf{compl}} \to \mathbb{M}^*_{\mathsf{S5}}$ is the bijection from Theorem 4.3.2, then for all $M = (X, P, V) \in \mathbb{M}_{\mathsf{int}} \cap \mathbb{M}_{\mathsf{compl}}$:

$$M, x \models \varphi \iff g(M), x \models t(\varphi) \tag{4.2}$$

Proof. Note that since g is defined as the restriction of f to $\mathbb{M}_{\mathsf{int}} \cap \mathbb{M}_{\mathsf{compl}}$, (4.2) follows from (4.1). We show (4.1) only. Let $M = (X, P, V) \in \mathbb{M}_{\mathsf{int}} \cap \mathbb{M}_{\mathsf{unions}}$. Write $f(M) = (X, R, V)$. From the definition of f and Lemma 4.3.2 (2), we have

$$A \in P \iff A \text{ is downwards closed wrt R} \qquad (*)$$

We show (4.1) by induction. The only non-trivial cases are E and S formulas.

- E: Suppose $M, x \models \mathsf{E}\varphi$. Then $\|\varphi\|_M \in P$. By the induction hypothesis and $(*)$, this means $\|t(\varphi)\|_{f(M)}$ is downwards closed with respect to R. Now take $y \in X$ such that $f(M), y \models \neg t(\varphi)$. Then $y \notin \|t(\varphi)\|_{f(M)}$. Since this set is downwards closed, it cannot contain any R-successor of y. Hence $f(M), y \models \mathsf{K}\neg t(\varphi)$. This shows that $f(M), x \models \mathsf{A}(\neg t(\varphi) \to \mathsf{K}\neg t(\varphi))$, i.e. $f(M), x \models t(\mathsf{E}\varphi)$.

  Now suppose $f(M), x \models t(\mathsf{E}\varphi)$, i.e. $f(M), x \models \mathsf{A}(\neg t(\varphi) \to \mathsf{K}\neg t(\varphi))$. We show $\|\varphi\|_M$ is downwards closed. Suppose yRz and $z \in \|\varphi\|_M$. By the induction hypothesis, $f(M), z \not\models \neg t(\varphi)$. Hence $f(M), y \not\models \mathsf{K}\neg t(\varphi)$. Since $\neg t(\varphi) \to \mathsf{K}\neg t(\varphi)$ holds everywhere in $f(M)$, this means $f(M), y \models t(\varphi)$; by the induction hypothesis again we get $M, y \models \varphi$ and thus $y \in \|\varphi\|_M$. This shows that $\|\varphi\|_M$ is downwards closed, and by $(*)$ we have $\|\varphi\|_M \in P$. Hence $M, x \models \mathsf{E}\varphi$.

- S: We show both directions by contraposition. Suppose $M, x \not\models \mathsf{S}\varphi$. Then there is $A \in P$ such that $\|\varphi\|_M \subseteq A$ and $x \notin A$. Since A is downwards closed (by $(*)$), this means xRy implies $y \notin A$ and hence $y \notin \|\varphi\|_M$, for any $y \in X$. By the induction hypothesis, we get that xRy implies $f(M), y \models \neg t(\varphi)$, i.e. $f(M), x \models \mathsf{K}\neg t(\varphi)$. Hence $f(M), x \not\models t(\mathsf{S}\varphi)$.

  Finally, suppose $f(M), x \not\models t(\mathsf{S}\varphi)$, i.e. $f(M), x \models \mathsf{K}\neg t(\varphi)$. Let A be the R-downwards closure of $\|\varphi\|_M$, i.e.

  $$A = \{y \in X \mid \exists z \in \|\varphi\|_M : yRz\}$$

  Then $\|\varphi\|_M \subseteq A$ by reflexivity of R, and A is downwards closed by transitivity. Hence $A \in P$. But $x \notin A$, since for all z with xRz we have $f(M), z \models \neg t(\varphi)$, so $z \notin \|t(\varphi)\|_{f(M)} = \|\varphi\|_M$. Hence $M, x \not\models \mathsf{S}\varphi$.

□

Taken together, the results of this section show that, when considering expertise models closed under intersections and unions, P *uniquely determines* an epistemic accessibility relation such that expertise and soundness have precise interpretations in terms of S4 knowledge. If we also impose closure under complements, the notion of knowledge is strengthened to S5. Moreover, every S4 and S5 model arises from some expertise model in this way.

These results also reflect back on the closure properties of expertise models. For example, since S5 is such a strong notion of knowledge – with the negative introspection axiom **5** not generally considered plausible for modelling human knowledge, for instance – we may conclude that closure of expertise under complements is a similarly strong assumption. Put differently, we may use the properties of knowledge corresponding to closure conditions to asses the desirability of the closure conditions themselves, making use of the extensive literature from the epistemic logic side in the process.

4.4 Axiomatisation

In this section we give sound and complete logics with respect to various classes of expertise models. We start with the class of all expertise models $\mathbb{M}$, and show how adding more axioms captures the closure conditions of Section 4.2.

Table 4.1: Axioms and inference rules for L.

$\mathsf{E}\varphi \leftrightarrow \mathsf{AE}\varphi$	(EA)
$\mathsf{A}(\varphi \leftrightarrow \psi) \rightarrow (\mathsf{E}\varphi \leftrightarrow \mathsf{E}\psi)$	(RE_{E})
$\mathsf{A}(\varphi \rightarrow \psi) \rightarrow (\mathsf{S}\varphi \wedge \mathsf{E}\psi \rightarrow \psi)$	(W_{E})
$\varphi \rightarrow \mathsf{S}\varphi$	(T_{S})
$\mathsf{SS}\varphi \rightarrow \mathsf{S}\varphi$	(4_{S})
$\mathsf{A}(\varphi \rightarrow \psi) \rightarrow (\mathsf{S}\varphi \rightarrow \mathsf{S}\psi)$	(W_{S})
$\mathsf{A}(\varphi \rightarrow \psi) \rightarrow (\mathsf{A}\varphi \rightarrow \mathsf{A}\psi)$	(K_{A})
$\mathsf{A}\varphi \rightarrow \varphi$	(T_{A})
$\neg\mathsf{A}\varphi \rightarrow \mathsf{A}\neg\mathsf{A}\varphi$	(5_{A})
From φ infer $\mathsf{A}\varphi$	($\mathrm{Nec}_{\mathsf{A}}$)
From $\varphi \rightarrow \psi$ and φ infer ψ	(MP)

The General Case. Let L be the extension of propositional logic generated by the axioms and inference rules shown in Table 4.1.[7] Note that we treat A as a "box" and S as a "diamond" modality. Some of the axioms were already seen in Proposition 4.1.1; new ones include "replacement of equivalents" for expertise (RE_{E}), **4** for S (4_{S}), and (W_{S}), which says that if ψ is logically weaker than φ then the same holds for $\mathsf{S}\psi$ and $\mathsf{S}\varphi$. First, L is sound, i.e. all formulas in L are valid.

Theorem 4.4.1. L *is sound with respect to* $\mathbb{M}$.

Proof. The inference rules clearly preserve validity. All axioms were either shown to be valid in Proposition 4.1.1 or are straightforward to see, with the possible exception of (4_{S}) which we will show explicitly. Let $M = (X, P, V)$ be an expertise model and $x \in X$. Suppose $M, x \models \mathsf{SS}\varphi$. We need to show $M, x \models \mathsf{S}\varphi$. Take $A \in P$ such that $\|\varphi\|_M \subseteq A$. Now for any $y \in X$, if $M, y \models \mathsf{S}\varphi$ then clearly $y \in A$. Hence $\|\mathsf{S}\varphi\|_M \subseteq A$. But then $M, x \models \mathsf{SS}\varphi$ gives $x \in A$. Hence $M, x \models \mathsf{S}\varphi$. ☐

For completeness we use a variation of the standard canonical model method (Blackburn, Rijke, and Venema 2001, §4.2). In taking this approach, one constructs a model whose states are maximally L-consistent sets of formulas, and aims to prove the *truth lemma*: that a set Γ satisfies φ in the canonical model if and only if $\varphi \in \Gamma$. However, the truth lemma poses some difficulties for our semantics. Roughly speaking, we find there is an obvious choice of P to ensure the truth lemma for $\mathsf{E}\varphi$ formulas, but that this may be too small for $\mathsf{S}\varphi$ to be refuted when $\mathsf{S}\varphi \notin \Gamma$ (recall that $M, x \not\models \mathsf{S}\varphi$ iff *there exists* some $A \in P$ such

[7]Formally, L is the smallest set of formulas which (i) contains all substitution instances of propositional tautologies (this means we include tautologies involving modalities, e.g. $\mathsf{E}p \vee \neg\mathsf{E}p$); (ii) contains all formulas taking the form of the axioms shown in Table 4.1; and (iii) is closed under the inference rules shown in Table 4.1.

that $\|\varphi\|_M \subseteq A$ and $x \notin A$). We therefore "enlargen" the set of states so we can add new expertise sets A – without affecting the truth value of expertise formulas – to obtain the truth lemma for soundness formulas.

First, some standard notation and terminology. Write $\vdash \varphi$ iff $\varphi \in \mathsf{L}$. For $\Gamma \subseteq \mathcal{L}$ and $\varphi \in \mathcal{L}$, write $\Gamma \vdash \varphi$ iff there are $\psi_0, \ldots, \psi_n \in \Gamma$, $n \geq 0$, such that $\vdash (\psi_0 \wedge \cdots \wedge \psi_n) \rightarrow \varphi$. Say Γ is *inconsistent* if $\Gamma \vdash \bot$, and *consistent* otherwise. Γ is *maximally consistent* iff Γ is consistent and $\Gamma \subset \Delta$ implies that Δ is inconsistent. We recall some standard facts about maximally consistent sets.

Lemma 4.4.1. *Let Γ be a maximally consistent set and $\varphi, \psi \in \mathcal{L}$. Then*

1. $\varphi \in \Gamma$ *iff* $\Gamma \vdash \varphi$
2. *If* $\varphi \rightarrow \psi \in \Gamma$ *and* $\varphi \in \Gamma$, *then* $\psi \in \Gamma$
3. $\neg\varphi \in \Gamma$ *iff* $\varphi \notin \Gamma$
4. $\varphi \wedge \psi \in \Gamma$ *iff* $\varphi \in \Gamma$ *and* $\psi \in \Gamma$

Proof.

1. First suppose $\varphi \in \Gamma$. Since $\varphi \rightarrow \varphi$ is an instance of the propositional tautology $p \rightarrow p$, we have $\vdash \varphi \rightarrow \varphi$. Since $\varphi \in \Gamma$, this gives $\Gamma \vdash \varphi$.

   Now suppose $\Gamma \vdash \varphi$. Set $\Delta = \Gamma \cup \{\varphi\}$. We claim Δ is consistent. If not, there are $\psi_0, \ldots, \psi_n \in \Delta$ such that $\vdash (\psi_0 \wedge \cdots \wedge \psi_n) \rightarrow \bot$. Since Γ is consistent, at least one of the ψ_i must be equal to φ. Without loss of generality, $\psi_0 = \varphi$ and $\psi_j \in \Gamma$ for $j > 0$. Hence, by propositional logic and (MP), $\vdash (\psi_1 \wedge \cdots \wedge \psi_n) \rightarrow \neg\varphi$. Thus $\Gamma \vdash \neg\varphi$. But since $\Gamma \vdash \varphi$ also, it follows that $\Gamma \vdash \bot$, and thus Γ is inconsistent: contradiction. So Δ must be consistent after all. Clearly $\Gamma \subseteq \Delta$, and by maximal consistency of Γ, $\Gamma \not\subset \Delta$. Hence $\Delta = \Gamma$, so $\varphi \in \Gamma$ as required.

2. By propositional logic we have $\vdash ((\varphi \rightarrow \psi) \wedge \varphi) \rightarrow \psi$. Hence $\Gamma \vdash \psi$; by (1) we get $\psi \in \Gamma$.

3. If $\neg\varphi \in \Gamma$ then clearly $\varphi \notin \Gamma$, since otherwise Γ would be inconsistent. If $\varphi \notin \Gamma$ then $\Gamma \nvdash \varphi$ by (1). Set $\Delta = \Gamma \cup \{\neg\varphi\}$. Then Δ is consistent (one can show that assuming Δ is inconsistent leads to $\Gamma \vdash \varphi$; a contradiction). Again, since $\Gamma \subseteq \Delta$ and Γ is maximally consistent, we must in fact have $\Gamma = \Delta$, so $\neg\varphi \in \Gamma$.

4. If $\varphi \wedge \psi \in \Gamma$ then both $\Gamma \vdash \varphi$ and $\Gamma \vdash \psi$, so $\varphi, \psi \in \Gamma$ by (1). Conversely, if $\varphi, \psi \in \Gamma$ then $\Gamma \vdash \varphi \wedge \psi$, so $\varphi \wedge \psi \in \Gamma$ by (1) again.

□

Lemma 4.4.2 (Lindenbaum's Lemma)**.** *If $\Gamma \subseteq \mathcal{L}$ is consistent there is a maximally consistent set Δ such that $\Gamma \subseteq \Delta$.*

Let X_{L} denote the set of maximally consistent sets. Define a relation R by

$$\Gamma R \Delta \iff \forall \varphi \in \mathcal{L} : \mathsf{A}\varphi \in \Gamma \implies \varphi \in \Delta$$

The (T_{A}) and (5_{A}) axioms for A show that R is an equivalence relation; this is part of the standard proof that S5 is complete for equivalence relations.

Lemma 4.4.3. *R is an equivalence relation.*

Proof. We first show that R is reflexive and has the *Euclidean property* (xRy and xRz implies yRz). For reflexivity, let $\Gamma \in X_{\mathsf{L}}$. Suppose $\mathsf{A}\varphi \in \Gamma$. By ($\mathrm{T}_{\mathsf{A}}$) and closure of maximally consistent sets under modus ponens, $\varphi \in \Gamma$. Hence $\Gamma R \Gamma$.

For the Euclidean property, suppose $\Gamma R \Delta$ and $\Gamma R \Lambda$. We show $\Delta R \Lambda$ by contraposition. Suppose $\varphi \notin \Lambda$. Since $\Gamma R \Lambda$, this means $\mathsf{A}\varphi \notin \Gamma$. Hence $\neg\mathsf{A}\varphi \in \Gamma$, and by ($5_{\mathsf{A}}$) we get $\mathsf{A}\neg\mathsf{A}\varphi \in \Gamma$. Now $\Gamma R \Delta$ gives $\neg\mathsf{A}\varphi \in \Delta$, so $\mathsf{A}\varphi \notin \Delta$.

To conclude we need to show R is symmetric and transitive. For symmetry, suppose $\Gamma R \Delta$. By reflexivity, $\Gamma R \Gamma$. The Euclidean property therefore gives $\Delta R \Gamma$. For transitivity, suppose $\Gamma R \Delta$ and $\Delta R \Lambda$. By symmetry, $\Delta R \Gamma$. The Euclidean property again gives $\Gamma R \Lambda$. □

For $\varphi \in \mathcal{L}$, let $|\varphi| = \{\Gamma \in X_{\mathsf{L}} \mid \varphi \in \Gamma\}$ be the *proof set* of φ. For $\Sigma \in X_{\mathsf{L}}$, let X_Σ be the equivalence class of Σ in R, and write $|\varphi|_\Sigma = |\varphi| \cap X_\Sigma$. Using what is essentially the standard proof of the truth lemma for the modal logic **K** with respect to relational semantics, (K_{A}) yields the following.

Lemma 4.4.4. *Let $\Sigma \in X_{\mathsf{L}}$. Then*

1. *For any $\varphi \in \mathcal{L}$, $\mathsf{A}\varphi \in \Sigma$ iff $|\varphi|_\Sigma = X_\Sigma$*
2. *For any $\varphi, \psi \in \mathcal{L}$, $\mathsf{A}(\varphi \to \psi) \in \Sigma$ iff $|\varphi|_\Sigma \subseteq |\psi|_\Sigma$*
3. *For any $\varphi, \psi \in \mathcal{L}$, $\mathsf{A}(\varphi \leftrightarrow \psi) \in \Sigma$ iff $|\varphi|_\Sigma = |\psi|_\Sigma$*

Proof.

1. For the left-to-right direction, suppose $\mathsf{A}\varphi \in \Sigma$. Let $\Gamma \in X_\Sigma$. Then $\Sigma R \Gamma$, so clearly $\varphi \in \Gamma$. Hence $|\varphi|_\Sigma = X_\Sigma$. For the other direction we show the contrapositive. Suppose $\mathsf{A}\varphi \notin \Sigma$. Set

$$\Gamma_0 = \{\psi \mid \mathsf{A}\psi \in \Gamma\} \cup \{\neg\varphi\}.$$

We claim Γ_0 is consistent. If not, without loss of generality there are $\psi_0, \ldots, \psi_n \in \Gamma_0$ such that $\mathsf{A}\psi_i \in \Sigma$ for each i, and $\vdash \psi_0 \wedge \cdots \wedge \psi_n \to \varphi$. By propositional logic, we get $\vdash \psi_0 \to \cdots \to \psi_n \to \varphi$ (where the implication arrows associate to the right) and by ($\mathrm{Nec}_{\mathsf{A}}$), $\vdash \mathsf{A}(\psi_0 \to \cdots \to \psi_n \to \varphi)$. Since ($\mathrm{K}_{\mathsf{A}}$) together with (MP) says that A distributes over implications, repeated applications gives $\vdash \mathsf{A}\psi_0 \to \cdots \to \mathsf{A}\psi_n \to \mathsf{A}\varphi$ and propositional logic again gives $\vdash \mathsf{A}\psi_0 \wedge \cdots \wedge \mathsf{A}\psi_n \to \mathsf{A}\varphi$. But recall that $\mathsf{A}\psi_i \in \Sigma$. Hence $\Sigma \vdash \mathsf{A}\varphi$. Since Σ is maximally consistent, this means $\mathsf{A}\varphi \in \Sigma$: contradiction.

So Γ_0 is consistent. By Lindenbaum's lemma (Lemma 4.4.2), there is a maximally consistent set $\Gamma \supseteq \Gamma_0$. Clearly $\Sigma R \Gamma$, since if $\mathsf{A}\psi \in \Sigma$ then $\psi \in \Gamma_0 \subseteq \Gamma$. Moreover, $\neg\varphi \in \Gamma_0 \subseteq \Gamma$, so by consistency $\varphi \notin \Gamma$. Hence $\Gamma \in X_\Sigma \setminus |\varphi|_\Sigma$, and we are done.

2. Note that by (1) we have

$$\begin{aligned} \mathsf{A}(\varphi \to \psi) \in \Sigma &\iff |\varphi \to \psi|_\Sigma = X_\Sigma \\ &\iff \forall \Gamma \in X_\Sigma : \varphi \to \psi \in \Gamma \end{aligned}$$

Suppose $\mathsf{A}(\varphi \to \psi) \in \Sigma$. Take $\Gamma \in |\varphi|_\Sigma$. Then we have $\varphi, \varphi \to \psi \in \Gamma$, so $\psi \in \Gamma$. This shows $|\varphi|_\Sigma \subseteq |\psi|_\Sigma$. Conversely, suppose $|\varphi|_\Sigma \subseteq |\psi|_\Sigma$. Take $\Gamma \in X_\Sigma$. If $\varphi \notin \Gamma$ then $\neg\varphi \in \Gamma$, so $\neg\varphi \vee \psi \in \Gamma$ and thus $\varphi \to \psi \in \Gamma$. If $\varphi \in \Gamma$ then $\Gamma \in |\varphi|_\Sigma \subseteq |\psi|_\Sigma$, so $\psi \in \Gamma$. Thus $\varphi \to \psi \in \Gamma$ in this case too. Hence $\mathsf{A}(\varphi \to \psi) \in \Sigma$.

3. First note that $\mathsf{A}(\alpha \wedge \beta) \in \Sigma$ iff both $\mathsf{A}\alpha \in \Sigma$ and $\mathsf{A}\beta \in \Sigma$. This can be shown using (K_A), (MP) and instances of the propositional tautologies $(p \wedge q) \to p$ (for the left-to-right implication) and $p \to q \to (p \wedge q))$ (for the right-to-left implication). Recalling that $\varphi \leftrightarrow \psi$ is an abbreviation for $(\varphi \to \psi) \wedge (\psi \to \varphi)$, we get

$$\begin{aligned} \mathsf{A}(\varphi \leftrightarrow \psi) \in \Sigma &\iff \mathsf{A}(\varphi \to \psi) \in \Sigma \text{ and } \mathsf{A}(\psi \to \varphi) \in \Sigma \\ &\iff |\varphi|_\Sigma \subseteq |\psi|_\Sigma \text{ and } |\psi|_\Sigma \subseteq |\varphi|_\Sigma \\ &\iff |\varphi|_\Sigma = |\psi|_\Sigma \end{aligned}$$

as required.

□

Corollary 4.4.1. *Let $\Sigma \in X_\mathsf{L}$. For $\Gamma, \Delta \in X_\Sigma$ and $\varphi \in \mathcal{L}$, $\mathsf{A}\varphi \in \Gamma$ iff $\mathsf{A}\varphi \in \Delta$ and $\mathsf{E}\varphi \in \Gamma$ iff $\mathsf{E}\varphi \in \Delta$.*

Proof. For the first point, note that if $\mathsf{A}\varphi \in \Gamma$ then Lemma 4.4.4 gives $|\varphi|_\Gamma = X_\Gamma$. But since Γ and Δ are in the same equivalence class of R, $|\varphi|_\Gamma = |\varphi|_\Delta$ and $X_\Gamma = X_\Delta$. Hence $|\varphi|_\Delta = X_\Delta$, so $\mathsf{A}\varphi \in \Delta$ by Lemma 4.4.4. The converse holds by symmetry.

For the second point, if $\mathsf{E}\varphi \in \Gamma$ then $\mathsf{AE}\varphi \in \Gamma$ by (EA). Since $\Gamma R \Delta$, we get $\mathsf{E}\varphi \in \Delta$. Again, the converse holds by symmetry. □

We are ready to define the "canonical" model (for each Σ). Set $\widehat{X}_\Sigma = X_\Sigma \times \mathbb{R}$. This is the step described informally above: we enlargen X_Σ by considering uncountably many copies of each point (any uncountable set would do in place of $\mathbb{R}$). The valuation is straightforward: set $\widehat{V}_\Sigma(p) = |p|_\Sigma \times \mathbb{R}$. For the expertise component of the model, say $A \subseteq \widehat{X}_\Sigma$ is *S-closed* iff for all $\varphi \in \mathcal{L}$:

$$|\varphi|_\Sigma \times \mathbb{R} \subseteq A \implies |\mathsf{S}\varphi|_\Sigma \times \mathbb{R} \subseteq A.$$

Set $\widehat{P}_\Sigma = \widehat{P}^0_\Sigma \cup \widehat{P}^1_\Sigma$, where

$$\widehat{P}^0_\Sigma = \{|\varphi|_\Sigma \times \mathbb{R} \mid \mathsf{E}\varphi \in \Sigma\},$$
$$\widehat{P}^1_\Sigma = \{A \subseteq \widehat{X}_\Sigma \mid A \text{ is S-closed and } \forall\varphi \in \mathcal{L} : A \neq |\varphi|_\Sigma \times \mathbb{R}\}.$$

We have a version of the truth lemma for the model $\widehat{M}_\Sigma = (\widehat{X}_\Sigma, \widehat{P}_\Sigma, \widehat{V}_\Sigma)$.

Lemma 4.4.5. *For any* $(\Gamma, t) \in \widehat{X}_\Sigma$ *and* $\varphi \in \mathcal{L}$,

$$\widehat{M}_\Sigma, (\Gamma, t) \models \varphi \iff \varphi \in \Gamma,$$

i.e. $\|\varphi\|_{\widehat{M}_\Sigma} = |\varphi|_\Sigma \times \mathbb{R}$.

Proof. By induction. The cases for atomic propositions and the propositional connectives are straightforward by the definition of $\widehat{V}_\Sigma$ and properties of maximally consistent sets. The case for the universal modality A is also straightforward by Lemma 4.4.4 and Corollary 4.4.1. We treat the cases of E and S formulas.

- E: First suppose $\mathsf{E}\varphi \in \Gamma$. By Corollary 4.4.1, $\mathsf{E}\varphi \in \Sigma$. Hence $|\varphi|_\Sigma \times \mathbb{R} \in \widehat{P}^0_\Sigma$. By the induction hypothesis, $\|\varphi\|_{\widehat{M}_\Sigma} \in \widehat{P}^0_\Sigma$. Hence $\widehat{M}_\Sigma, (\Gamma, t) \models \mathsf{E}\varphi$.

  Now suppose $\widehat{M}_\Sigma, (\Gamma, t) \models \mathsf{E}\varphi$. Then $\|\varphi\|_{\widehat{M}_\Sigma} \in \widehat{P}_\Sigma$. By the induction hypothesis, $\|\varphi\|_{\widehat{M}_\Sigma} = |\varphi|_\Sigma \times \mathbb{R}$. Hence $|\varphi|_\Sigma \times \mathbb{R} \in \widehat{P}_\Sigma$. Since $\widehat{P}^1_\Sigma$ does not contain any sets of this form, we must have $|\varphi|_\Sigma \times \mathbb{R} \in \widehat{P}^0_\Sigma$. Therefore there is some ψ such that $\mathsf{E}\psi \in \Sigma$ and $|\varphi|_\Sigma \times \mathbb{R} = |\psi|_\Sigma \times \mathbb{R}$. It follows that $|\varphi|_\Sigma = |\psi|_\Sigma$, and Lemma 4.4.4 then gives $\mathsf{A}(\varphi \leftrightarrow \psi) \in \Sigma$. By Corollary 4.4.1, we have $\mathsf{E}\psi \in \Gamma$ and $\mathsf{A}(\varphi \leftrightarrow \psi) \in \Gamma$ too. By (RE_E) we get $\mathsf{E}\varphi \in \Gamma$ as required.

- S: First suppose $\mathsf{S}\varphi \in \Gamma$. Take $A \in \widehat{P}_\Sigma$ such that $\|\varphi\|_{\widehat{M}_\Sigma} \subseteq A$. By the induction hypothesis, $|\varphi|_\Sigma \times \mathbb{R} \subseteq A$. There are two cases: either $A \in \widehat{P}^0_\Sigma$ or $A \in \widehat{P}^1_\Sigma$.

  If $A \in \widehat{P}^0_\Sigma$, there is ψ such that $A = |\psi|_\Sigma \times \mathbb{R}$ and $\mathsf{E}\psi \in \Sigma$. Since $|\varphi|_\Sigma \times \mathbb{R} \subseteq A$, we have $|\varphi|_\Sigma \subseteq |\psi|_\Sigma$. By Lemma 4.4.4, $\mathsf{A}(\varphi \to \psi) \in \Sigma$. By Corollary 4.4.1 we have $\mathsf{E}\psi, \mathsf{A}(\varphi \to \psi) \in \Gamma$ too. Applying (W_E) gives $\mathsf{S}\varphi \wedge \mathsf{E}\psi \to \psi \in \Gamma$; since $\mathsf{S}\varphi, \mathsf{E}\psi \in \Gamma$ we have $\mathsf{S}\varphi \wedge \mathsf{E}\psi \in \Gamma$ and thus $\psi \in \Gamma$. This means $(\Gamma, t) \in |\psi|_\Sigma \times \mathbb{R} = A$, as required.

  If $A \in \widehat{P}^1_\Sigma$, A is S-closed by definition. Hence $|\mathsf{S}\varphi|_\Sigma \times \mathbb{R} \subseteq A$. Since $\mathsf{S}\varphi \in \Gamma$ we get $(\Gamma, t) \in A$ as required. In either case we have $(\Gamma, t) \in A$. This shows $\widehat{M}_\Sigma, (\Gamma, t) \models \mathsf{S}\varphi$.

For the other direction we show the contrapositive. Take any $(\Gamma, t) \in \widehat{X}_\Sigma$ and suppose $\mathsf{S}\varphi \notin \Gamma$. We show that $\widehat{M}_\Sigma, (\Gamma, t) \not\models \mathsf{S}\varphi$, i.e. there is $A \in \widehat{P}_\Sigma$ such that $\|\varphi\|_{\widehat{M}_\Sigma} \subseteq A$ but $(\Gamma, t) \notin A$. First, set

$$\mathcal{U} = \{|\psi|_\Sigma \times \mathbb{R} \mid \psi \in \mathcal{L} \text{ and } |\psi|_\Sigma \times \mathbb{R} \not\subseteq |\mathsf{S}\varphi|_\Sigma \times \mathbb{R}\}.$$

Since $\mathcal{L}$ is countable, $\mathcal{U}$ is at most countable. Hence we may write $\mathcal{U} = \{U_n\}_{n \in N}$ for some index set $N \subseteq \mathbb{N}$. Since $U_n \not\subseteq |\mathsf{S}\varphi|_\Sigma \times \mathbb{R}$, we may choose some $(\Delta_n, t_n) \in U_n \setminus (|\mathsf{S}\varphi|_\Sigma \times \mathbb{R})$ for each n. Now write

$$\mathcal{D} = \{(\Delta_n, t_n)\}_{n \in N} \cup \{(\Gamma, t)\}.$$

Since N is at most countable, so is $\mathcal{D}$. Since $\mathbb{R}$ is uncountable, there is some $s \in \mathbb{R}$ such that $(\Gamma, s) \notin \mathcal{D}$.[8] We necessarily have $s \neq t$. We are ready to define A: set

$$A = (|\mathsf{S}\varphi|_\Sigma \times \mathbb{R}) \cup \{(\Gamma, s)\}.$$

Note that $(\Gamma, t) \notin A$ since $\mathsf{S}\varphi \notin \Gamma$ and $s \neq t$. Next we show $\|\varphi\|_{\widehat{M}_\Sigma} \subseteq A$. By the induction hypothesis, this is equivalent to $|\varphi|_\Sigma \times \mathbb{R} \subseteq A$. By ($\mathrm{T}_\mathsf{S}$) and ($\mathrm{Nec}_\mathsf{A}$), we have $\mathsf{A}(\varphi \to \mathsf{S}\varphi) \in \Sigma$, and consequently $|\varphi|_\Sigma \subseteq |\mathsf{S}\varphi|_\Sigma$ by Lemma 4.4.4. Hence $|\varphi|_\Sigma \times \mathbb{R} \subseteq |\mathsf{S}\varphi|_\Sigma \times \mathbb{R} \subseteq A$ as required.

It only remains to show that $A \in \widehat{P}_\Sigma$. We claim that $A \in \widehat{P}^1_\Sigma$. First, A is S-closed. Indeed, suppose $|\psi|_\Sigma \times \mathbb{R} \subseteq A$. We claim that, in fact, $|\psi|_\Sigma \times \mathbb{R} \subseteq |\mathsf{S}\varphi|_\Sigma \times \mathbb{R}$. If not, then by definition of $\mathcal{U}$ there is $n \in N$ such that $|\psi|_\Sigma \times \mathbb{R} = U_n$. Hence $U_n \subseteq A$. This means $(\Delta_n, t_n) \in A$. But $(\Delta_n, t_n) \notin |\mathsf{S}\varphi|_\Sigma \times \mathbb{R}$, so we must have $(\Delta_n, t_n) = (\Gamma, s)$. But then $(\Gamma, s) \in \mathcal{D}$: contradiction. So we do indeed have $|\psi|_\Sigma \times \mathbb{R} \subseteq |\mathsf{S}\varphi|_\Sigma \times \mathbb{R}$, and thus $|\psi|_\Sigma \subseteq |\mathsf{S}\varphi|_\Sigma$. By Lemma 4.4.4, $\mathsf{A}(\psi \to \mathsf{S}\varphi) \in \Sigma$.

Now, take any $(\Lambda, u) \in |\mathsf{S}\psi|_\Sigma \times \mathbb{R}$. Since $\Lambda \in X_\Sigma$, Corollary 4.4.1 gives $\mathsf{A}(\psi \to \mathsf{S}\varphi) \in \Lambda$. By ($\mathrm{W}_\mathsf{S}$), $\mathsf{S}\psi \to \mathsf{SS}\varphi \in \Lambda$. Since $\Lambda \in |\mathsf{S}\psi|_\Sigma$, we get $\mathsf{SS}\varphi \in \Lambda$. But then ($4_\mathsf{S}$) gives $\mathsf{S}\varphi \in \Lambda$. That is, $(\Lambda, u) \in |\mathsf{S}\varphi|_\Sigma \times \mathbb{R} \subseteq A$. This shows $|\mathsf{S}\psi|_\Sigma \times \mathbb{R} \subseteq A$, so A is S-closed.

Finally, we show that for all $\psi \in \mathcal{L}$, $A \neq |\psi|_\Sigma \times \mathbb{R}$. For contradiction, suppose there is ψ with $A = |\psi|_\Sigma \times \mathbb{R}$. Then since $(\Gamma, s) \in A$, we have $\Gamma \in |\psi|_\Sigma$. But then $(\Gamma, t) \in |\psi|_\Sigma \times \mathbb{R} = A$: contradiction.

This completes the proof that $A \in \widehat{P}^1_\Sigma$. Thus $\widehat{M}_\Sigma, (\Gamma, t) \not\models \mathsf{S}\varphi$, and we are done.

□

We are finally in a position to show completness. In fact, we have *strong completeness*: for all sets $\Gamma \subseteq \mathcal{L}$ and $\varphi \in \mathcal{L}$, if $\Gamma \models \varphi$ then $\Gamma \vdash \varphi$.

[8] If not, then $s \mapsto (\Gamma, s)$ is an injective mapping $\mathbb{R} \to \mathcal{D}$, which would imply $\mathbb{R}$ is countable.

Theorem 4.4.2. L *is strongly complete with respect to* $\mathbb{M}$.

Proof. We show the contrapositive. Suppose $\Gamma \nvdash \varphi$. Then $\Gamma \cup \{\neg\varphi\}$ is consistent. By Lindenbaum's Lemma, there is a maximally consistent set $\Sigma \supseteq \Gamma \cup \{\neg\varphi\}$. Consider the model $\widehat{M_\Sigma}$. For any $\psi \in \Gamma$ we have $\psi \in \Sigma$, so Lemma 4.4.5 (with $t = 0$, say) gives $\widehat{M_\Sigma}, (\Sigma, 0) \models \psi$. Also, $\neg\varphi \in \Gamma \subseteq \Sigma$ gives $\widehat{M_\Sigma}, (\Sigma, 0) \models \neg\varphi$, so $\widehat{M_\Sigma}, (\Sigma, 0) \not\models \varphi$. This shows that $\Gamma \not\models \varphi$, and we are done. □

From soundness, which was shown in Theorem 4.4.1, one can easily show that $\Gamma \vdash \varphi$ implies $\Gamma \models \varphi$. Together with strong completeness, we get $\Gamma \models \varphi$ if and only if $\Gamma \vdash \varphi$. That is, semantic entailment with respect to the class of all models $\mathbb{M}$ is exactly the same notion as syntactic entailment using the axioms and inference rules of L.

Extensions of the Base Logic. We now extend L to obtain axiomatisations of sub-classes of $\mathbb{M}$ corresponding to closure conditions.

To start, consider closure under intersections. It was shown in Proposition 4.2.1 that the formula $\mathsf{A}(\mathsf{S}\varphi \rightarrow \varphi) \rightarrow \mathsf{E}\varphi$ characterises frames closed under intersections. It is perhaps no surprise that adding this as an axiom results in a sound and complete axiomatisation of $\mathbb{M}_{\mathsf{int}}$. Formally, let $\mathsf{L}_{\mathsf{int}}$ be the extension of L with the following axiom

$$\mathsf{A}(\mathsf{S}\varphi \rightarrow \varphi) \rightarrow \mathsf{E}\varphi \quad (\mathrm{Red}_{\mathsf{E}}),$$

so-named since together with $\mathsf{E}\varphi \rightarrow \mathsf{A}(\mathsf{S}\varphi \rightarrow \varphi)$ – which is derivable in L – it allows expertise to be reduced to soundness. That is, expertise on φ is equivalent to the statement that, in all situations, φ is only true up to lack of expertise if it is in fact true.

Theorem 4.4.3. $\mathsf{L}_{\mathsf{int}}$ *is sound and strongly complete with respect to* $\mathbb{M}_{\mathsf{int}}$.

Proof. For soundness, we only need to check that $(\mathrm{Red}_{\mathsf{E}})$ is sound for $\mathbb{M}_{\mathsf{int}}$. But this follows from Proposition 4.2.1 (1).

For completeness, we adopt a roughly similar approach to the general case. Let consistency, maximal consistency and other standard notions and notation be defined as before, but now for $\mathsf{L}_{\mathsf{int}}$ instead of L. Let $X_{\mathsf{L}_{\mathsf{int}}}$ be the set of maximally $\mathsf{L}_{\mathsf{int}}$-consistent sets. Define the relation R on $X_{\mathsf{L}_{\mathsf{int}}}$ in exactly the same way. Since $\mathsf{L}_{\mathsf{int}}$ extends L, R is again an equivalence relation, and we have the analogues of Lemma 4.4.4 and Corollary 4.4.1.

This time, however, the construction of the canonical model for a given $\Sigma \in X_{\mathsf{L}_{\mathsf{int}}}$ is much more straightforward. The set of states is simply X_Σ, i.e. the equivalence class of Σ in R. Overriding earlier terminology, say $A \subseteq X_\Sigma$ is *S-closed* iff $|\varphi|_\Sigma \subseteq A$ implies $|\mathsf{S}\varphi|_\Sigma \subseteq A$ for all $\varphi \in \mathcal{L}$. Then set

$$P_\Sigma = \{A \subseteq X_\Sigma \mid A \text{ is S-closed}\}.$$

Finally, set $V_\Sigma(p) = |p|_\Sigma$, and write $M_\Sigma = (X_\Sigma, P_\Sigma, V_\Sigma)$.

First, we have $M_\Sigma \in \mathbb{M}_{\mathsf{int}}$, i.e. intersections of S-closed sets are S-closed. Indeed, suppose $\{A_i\}_{i \in I}$ is a collection of S-closed sets, and suppose $|\varphi|_\Sigma \subseteq \bigcap_{i \in I} A_i$. Then $|\varphi|_\Sigma \subseteq A_i$ for each i, so S-closure gives $|\mathsf{S}\varphi|_\Sigma \subseteq A_i$. Hence $|\mathsf{S}\varphi|_\Sigma \subseteq \bigcap_{i \in I} A_i$.

Importantly, we have the truth lemma for M_Σ: for all $\Gamma \in X_\Sigma$ and $\varphi \in \mathcal{L}$,

$$M_\Sigma, \Gamma \models \varphi \iff \varphi \in \Gamma,$$

i.e. $\|\varphi\|_{M_\Sigma} = |\varphi|_\Sigma$.

As usual, the proof is by induction on formulas. The case for atomic propositions follows from the definition of V_Σ, the cases for conjunctions and negations hold by properties of maximally consistent sets, and the case for $\mathsf{A}\varphi$ holds by an argument identical to the one used in the general case (Lemma 4.4.5). The only interesting cases are therefore for $\mathsf{E}\varphi$ and $\mathsf{S}\varphi$ formulas:

- E: First suppose $\mathsf{E}\varphi \in \Gamma$. We claim $|\varphi|_\Sigma$ is S-closed. This will give $\|\varphi\|_{M_\Sigma} \in P_\Sigma$ by the induction hypothesis and definition of P_Σ, and therefore $M_\Sigma, \Gamma \models \mathsf{E}\varphi$.

  So, suppose $|\psi|_\Sigma \subseteq |\varphi|_\Sigma$. Then $\mathsf{A}(\psi \to \varphi) \in \Sigma$. Let $\Delta \in |\mathsf{S}\psi|_\Sigma$. Since $\Delta, \Gamma, \Sigma \in X_\Sigma$, we have $\mathsf{E}\varphi \in \Delta$ and $\mathsf{A}(\psi \to \varphi) \in \Delta$ too. By (W_{E}), $\mathsf{S}\psi \wedge \mathsf{E}\varphi \to \varphi \in \Delta$. But $\mathsf{S}\psi \in \Delta$, so $\mathsf{S}\psi \wedge \mathsf{E}\varphi \in \Delta$ and thus $\varphi \in \Delta$, i.e. $\Delta \in |\varphi|_\Sigma$. This shows $|\mathsf{S}\psi|_\Sigma \subseteq |\varphi|_\Sigma$, so $|\varphi|_\Sigma$ is S-closed as required.

  Now suppose $M_\Sigma, \Gamma \models \mathsf{E}\varphi$. Then, by the induction hypothesis, $|\varphi|_\Sigma$ is S-closed. Since $|\varphi|_\Sigma \subseteq |\varphi|_\Sigma$ clearly holds, we get $|\mathsf{S}\varphi|_\Sigma \subseteq |\varphi|_\Sigma$. This implies $\mathsf{A}(\mathsf{S}\varphi \to \varphi) \in \Sigma$, and ($\mathrm{Red}_{\mathsf{E}}$) gives $\mathsf{E}\varphi \in \Sigma$. Since $\Gamma \in X_\Sigma$, we get $\mathsf{E}\varphi \in \Gamma$ as required.

- S: Suppose $\mathsf{S}\varphi \in \Gamma$. Take any $A \in P_\Sigma$ such that $\|\varphi\|_{M_\Sigma} \subseteq A$. By the induction hypothesis, $|\varphi|_\Sigma \subseteq A$. By S-closure of A, $|\mathsf{S}\varphi|_\Sigma \subseteq A$. Hence $\Gamma \in |\mathsf{S}\varphi|_\Sigma \subseteq A$. This shows $M_\Sigma, \Gamma \models \mathsf{S}\varphi$.

  For the other direction we show the contrapositive. Suppose $\mathsf{S}\varphi \notin \Gamma$. First, we claim $|\mathsf{S}\varphi|_\Sigma$ is S-closed. Indeed, suppose $|\psi|_\Sigma \subseteq |\mathsf{S}\varphi|_\Sigma$. Then $\mathsf{A}(\psi \to \mathsf{S}\varphi) \in \Sigma$. Take any $\Delta \in |\mathsf{S}\psi|_\Sigma$. Since $\Delta \in X_\Sigma$, $\mathsf{A}(\psi \to \mathsf{S}\varphi) \in \Delta$ also. By (W_{S}), $\mathsf{S}\psi \to \mathsf{S}\mathsf{S}\varphi \in \Delta$. Now $\mathsf{S}\psi \in \Delta$ implies $\mathsf{S}\mathsf{S}\varphi \in \Delta$, and ($4_{\mathsf{S}}$) gives $\mathsf{S}\varphi \in \Delta$, i.e. $\Delta \in |\mathsf{S}\varphi|_\Sigma$. This shows $|\mathsf{S}\psi|_\Sigma \subseteq |\mathsf{S}\varphi|_\Sigma$, and thus $|\mathsf{S}\varphi|_\Sigma$ is S-closed.

  Hence $|\mathsf{S}\varphi|_\Sigma$ is a set in P_Σ not containing Γ. Moreover, $\|\varphi\|_{M_\Sigma} \subseteq |\mathsf{S}\varphi|_\Sigma$ by the induction hypothesis and (T_{S}). Hence $M_\Sigma, \Gamma \not\models \mathsf{S}\varphi$.

Strong completeness now follows. If $\Gamma \nvdash_{\mathsf{L}_{\mathsf{int}}} \varphi$, then $\Gamma \cup \{\neg\varphi\}$ is consistent, so by Lindenbaum's Lemma there is $\Sigma \in X_{\mathsf{L}_{\mathsf{int}}}$ with $\Sigma \supseteq \Gamma \cup \{\neg\varphi\}$. Considering the model $M_\Sigma \in \mathbb{M}_{\mathsf{int}}$, we have $M_\Sigma, \Sigma \models \Gamma$ and $M_\Sigma, \Sigma \not\models \varphi$ by the truth lemma. Hence $\Gamma \not\models_{\mathbb{M}_{\mathsf{int}}} \varphi$. □

Now we add finite unions to the mix. It was shown in Proposition 4.2.2 that within class $\mathbb{M}_{\mathsf{int}}$, the **K** axiom for the dual operator $\hat{\mathsf{S}}\varphi = \neg\mathsf{S}\neg\varphi$ characterises closure under finite unions. Note that any frame (X, P) closed under intersections and finite unions is a topological space,[9] where P is the set of *closed* sets. Write $\mathbb{M}_{\mathsf{top}} = \mathbb{M}_{\mathsf{int}} \cap \mathbb{M}_{\mathsf{finite-unions}}$ for the class of models over such frames. We obtain an axiomatisation of $\mathbb{M}_{\mathsf{top}}$ by adding **K** for $\hat{\mathsf{S}}$ and a bridge axiom linking $\hat{\mathsf{S}}$ and A:

$$\begin{array}{ll} \hat{\mathsf{S}}(\varphi \to \psi) \to (\hat{\mathsf{S}}\varphi \to \hat{\mathsf{S}}\psi) & (\mathrm{K}_{\mathsf{S}}) \\ \mathsf{A}\varphi \to \hat{\mathsf{S}}\varphi & (\mathrm{Inc}) \end{array}$$

Let $\mathsf{L}_{\mathsf{top}}$ be the extension of $\mathsf{L}_{\mathsf{int}}$ by $(\mathrm{K}_{\mathsf{S}})$ and (Inc). Note that $\mathsf{L}_{\mathsf{top}}$ contains the **KT4** axioms for $\hat{\mathsf{S}}$ (recalling that $(\mathrm{T}_{\mathsf{S}})$ and (4_{S}) are the "diamond" versions of **T** and **4**). Since **KT4** together with the bridge axiom (Inc) is complete for the class of relational models $\mathbb{M}^*_{\mathsf{S4}}$, we can exploit Theorem 4.3.3 to obtain completeness of $\mathsf{L}_{\mathsf{top}}$ with respect to $\mathbb{M}_{\mathsf{int}} \cap \mathbb{M}_{\mathsf{unions}}$. Since this class is included in $\mathbb{M}_{\mathsf{top}}$, we also get completeness with respect to $\mathbb{M}_{\mathsf{top}}$.[10]

Theorem 4.4.4. $\mathsf{L}_{\mathsf{top}}$ *is sound and strongly complete with respect to* $\mathbb{M}_{\mathsf{top}}$.

Proof. Soundness of $(\mathrm{K}_{\mathsf{S}})$ for $\mathbb{M}_{\mathsf{top}}$ follows from Proposition 4.2.2. For (Inc), suppose $M = (X, P, V) \in \mathbb{M}_{\mathsf{top}}$, $x \in X$ and $M, x \models \mathsf{A}\varphi$. Then $\|\varphi\|_M = X$, so $\|\neg\varphi\|_M = \emptyset$. By the convention that the empty set is the empty union $\bigcup\emptyset$ (which is a finite union), we have $\emptyset \in P$. Taking $A = \emptyset$ in the definition of the semantics for S, we have $\|\neg\varphi\|_M \subseteq A$ but clearly $x \notin A$. Hence $M, x \not\models \mathsf{S}\neg\varphi$, so $M, x \models \hat{\mathsf{S}}\varphi$.

For completeness, we go via relational semantics using the translation $t : \mathcal{L} \to \mathcal{L}_{\mathsf{KA}}$ and Theorem 4.3.3. First, let $\mathsf{L}_{\mathsf{S4A}}$ be the logic of $\mathcal{L}_{\mathsf{KA}}$ formulas formed by the axioms and inference rules shown in Table 4.2. It is well known that $\mathsf{L}_{\mathsf{S4A}}$ is strongly complete with respect to $\mathbb{M}^*_{\mathsf{S4}}$ (Blackburn, Rijke, and Venema 2001, Theorem 7.2).

Now, define a translation $u : \mathcal{L}_{\mathsf{KA}} \to \mathcal{L}$ as follows:

$$\begin{array}{ll} u(p) & = p \\ u(\varphi \wedge \psi) & = u(\varphi) \wedge u(\psi) \\ u(\neg\varphi) & = \neg u(\varphi) \\ u(\mathsf{K}\varphi) & = \neg\mathsf{S}\neg u(\varphi) \\ u(\mathsf{A}\varphi) & = \mathsf{A}u(\varphi). \end{array}$$

Recall the translation $t : \mathcal{L} \to \mathcal{L}_{\mathsf{KA}}$ from Section 4.3. While u is not the inverse of t (for instance, there is no $\psi \in \mathcal{L}_{\mathsf{KA}}$ with $u(\psi) = \mathsf{E}p$), for any $\varphi \in \mathcal{L}$ we have that φ is $\mathsf{L}_{\mathsf{top}}$-provably equivalent to $u(t(\varphi))$.

[9] By the convention that the empty intersection is the whole space X and the empty union is $\emptyset$, we have $X, \emptyset \in P$ too.

[10] Note that **KT4** is also complete for topological spaces with respect to the interior semantics (van Benthem and Bezhanishvili 2007).

Table 4.2: Axioms and inference rules for $\mathsf{L}_{\mathsf{S4A}}$.

$\mathsf{K}(\varphi \to \psi) \to (\mathsf{K}\varphi \to \mathsf{K}\psi)$	(K_{K})
$\mathsf{K}\varphi \to \varphi$	(T_{K})
$\mathsf{K}\varphi \to \mathsf{K}\mathsf{K}\varphi$	(4_{K})
$\mathsf{A}(\varphi \to \psi) \to (\mathsf{A}\varphi \to \mathsf{A}\psi)$	(K_{A})
$\mathsf{A}\varphi \to \varphi$	(T_{A})
$\neg\mathsf{A}\varphi \to \mathsf{A}\neg\mathsf{A}\varphi$	(5_{A})
$\mathsf{A}\varphi \to \mathsf{K}\varphi$	($\mathrm{Inc}_{\mathsf{K}}$)
From φ infer $\mathsf{A}\varphi$	($\mathrm{Nec}_{\mathsf{A}}$)
From $\varphi \to \psi$ and φ infer ψ	(MP)

Claim 4.4.1. *Let* $\varphi \in \mathcal{L}$. *Then* $\vdash_{\mathsf{L}_{\mathsf{top}}} \varphi \leftrightarrow u(t(\varphi))$.

Proof of claim. By induction on $\mathcal{L}$ formulas. The cases of atomic propositions and propositional connectives are straightforward. For the other cases, first note that the "replacement of equivalents" rule is derivable in L (and thus in $\mathsf{L}_{\mathsf{top}}$) for S, E and A:

$$\text{From } \varphi \leftrightarrow \psi \text{ infer } \bigcirc\varphi \leftrightarrow \bigcirc\psi \quad (\bigcirc \in \{\mathsf{S}, \mathsf{E}, \mathsf{A}\}).$$

For S this follows from ($\mathrm{Nec}_{\mathsf{A}}$) and ($\mathrm{W}_{\mathsf{S}}$); for E from ($\mathrm{Nec}_{\mathsf{A}}$) and ($\mathrm{RE}_{\mathsf{E}}$), and for A from ($\mathrm{Nec}_{\mathsf{A}}$) and ($\mathrm{K}_{\mathsf{A}}$). Now for the inductive step, suppose $\vdash_{\mathsf{L}_{\mathsf{top}}} \varphi \leftrightarrow u(t(\varphi))$.

- S: Note that

$$u(t(\mathsf{S}\varphi)) = u(\neg\mathsf{K}\neg t(\varphi)) = \neg\neg\mathsf{S}\neg\neg u(t(\varphi)).$$

  By the inductive hypothesis, propositional logic and replacement of equivalents, $\vdash_{\mathsf{L}_{\mathsf{top}}} \mathsf{S}\varphi \leftrightarrow u(t(\mathsf{S}\varphi))$.

- E: We have

$$\begin{aligned} u(t(\mathsf{E}\varphi)) &= u(\mathsf{A}(\neg t(\varphi) \to \mathsf{K}\neg t(\varphi))) \\ &= \mathsf{A}u(\neg t(\varphi) \to \mathsf{K}\neg t(\varphi)) \\ &= \mathsf{A}(u(\neg t(\varphi)) \to u(\mathsf{K}\neg t(\varphi))) \\ &= \mathsf{A}(\neg u(t(\varphi)) \to \neg\mathsf{S}\neg u(\neg t(\varphi))) \\ &= \mathsf{A}(\neg u(t(\varphi)) \to \neg\mathsf{S}\neg\neg u(t(\varphi))). \end{aligned}$$

  Taking the contrapositive of the implication, and using replacement of equivalents together with the inductive hypothesis, we get

$$\vdash_{\mathsf{L}_{\mathsf{top}}} u(t(\mathsf{E}\varphi)) \leftrightarrow \mathsf{A}(\mathsf{S}\varphi \to \varphi).$$

  But we have already seen that $\vdash_{\mathsf{L}_{\mathsf{int}}} \mathsf{E}\varphi \leftrightarrow \mathsf{A}(\mathsf{S}\varphi \to \varphi)$; since $\mathsf{L}_{\mathsf{top}}$ extends $\mathsf{L}_{\mathsf{int}}$, we get $\vdash_{\mathsf{L}_{\mathsf{top}}} \mathsf{E}\varphi \leftrightarrow u(t(\mathsf{E}\varphi))$.

- A: This case is straightforward by the inductive hypothesis and replacement of equivalents, since $u(t(\mathsf{A}\varphi)) = \mathsf{A}u(t(\varphi))$.

$\diamondsuit$

Next we show that if $\varphi \in \mathcal{L}_{\mathsf{KA}}$ is a theorem of $\mathsf{L}_{\mathsf{S4A}}$, then $u(\varphi)$ is a theorem of $\mathsf{L}_{\mathsf{top}}$.

Claim 4.4.2. *Let $\varphi \in \mathcal{L}_{\mathsf{KA}}$. Then $\vdash_{\mathsf{L}_{\mathsf{S4A}}} \varphi$ implies $\vdash_{\mathsf{L}_{\mathsf{top}}} u(\varphi)$.*

Proof of claim. By induction on the length of $\mathsf{L}_{\mathsf{S4A}}$ proofs. The base case consists of showing that if φ is an instance of an $\mathsf{L}_{\mathsf{S4A}}$ axiom or a substitution instance of a propositional tautology, then $\vdash_{\mathsf{L}_{\mathsf{top}}} u(\varphi)$. The case for instances of tautologies is straightforward, since u does not affect the structure of a propositional formula. We take the axioms of $\mathsf{L}_{\mathsf{S4A}}$ in turn.

- (K_{K}): We have

$$\begin{aligned} &u(\mathsf{K}(\varphi \to \psi) \to (\mathsf{K}\varphi \to \mathsf{K}\psi)) \\ &\quad = \neg\mathsf{S}\neg(u(\varphi) \to u(\psi)) \to (\neg\mathsf{S}\neg u(\varphi) \to \neg\mathsf{S}\neg u(\psi)) \\ &\quad = \hat{\mathsf{S}}(u(\varphi) \to u(\psi)) \to (\hat{\mathsf{S}}u(\varphi) \to \hat{\mathsf{S}}u(\psi)) \end{aligned}$$

  which is an instance of (K_{S}).

- (T_{K}): We have
$$u(\mathsf{K}\varphi \to \varphi) = \neg\mathsf{S}\neg u(\varphi) \to u(\varphi)$$
  Taking the contrapositive, this is $\mathsf{L}_{\mathsf{top}}$-provably equivalent to $\neg u(\varphi) \to \mathsf{S}\neg u(\varphi)$, which is an instance of (T_{S}).

- (4_{K}): We have

$$u(\mathsf{K}\varphi \to \mathsf{K}\mathsf{K}\varphi) = \neg\mathsf{S}\neg u(\varphi) \to \neg\mathsf{S}\neg\neg\mathsf{S}\neg u(\varphi)$$

  This is provably equivalent to $\mathsf{S}\mathsf{S}\neg u(\varphi) \to \mathsf{S}\neg u(\varphi)$, which is an instance of (4_{S}).

- (K_{A}): We have

$$u(\mathsf{A}(\varphi \to \psi) \to (\mathsf{A}\varphi \to \mathsf{A}\psi)) = \mathsf{A}(u(\varphi) \to u(\psi)) \to (\mathsf{A}u(\varphi) \to \mathsf{A}u(\psi))$$

  which is an instance of (K_{A}) in $\mathsf{L}_{\mathsf{top}}$.

- (T_{A}): We have
$$u(\mathsf{A}\varphi \to \varphi) = \mathsf{A}u(\varphi) \to u(\varphi)$$
  which is an instance of (T_{A}) in $\mathsf{L}_{\mathsf{top}}$.

- (5_{A}): We have

$$u(\neg\mathsf{A}\varphi \to \mathsf{A}\neg\mathsf{A}\varphi) = \neg\mathsf{A}u(\varphi) \to \mathsf{A}\neg\mathsf{A}u(\varphi)$$

  which is an instance of (5_{A}) in $\mathsf{L}_{\mathsf{top}}$.

- $(\mathrm{Inc}_{\mathsf{K}})$: We have

$$u(\mathsf{A}\varphi \to \mathsf{K}\varphi) = \mathsf{A}u(\varphi) \to \neg\mathsf{S}\neg u(\varphi) = \mathsf{A}u(\varphi) \to \hat{\mathsf{S}}u(\varphi)$$

  which is an instance of (Inc).

For the inductive step, we show that for each inference rule $\frac{\psi_1,\dots,\psi_n}{\varphi}$, if $\vdash_{\mathsf{L}_{\mathsf{top}}} u(\psi_i)$ for each i then $\vdash_{\mathsf{L}_{\mathsf{top}}} u(\varphi)$.

- $(\mathrm{Nec}_{\mathsf{A}})$: If $\vdash_{\mathsf{L}_{\mathsf{top}}} u(\varphi)$, then from $(\mathrm{Nec}_{\mathsf{A}})$ in $\mathsf{L}_{\mathsf{top}}$ we get $\vdash_{\mathsf{L}_{\mathsf{top}}} \mathsf{A}u(\varphi)$. But $\mathsf{A}u(\varphi) = u(\mathsf{A}\varphi)$, so we are done.

- (MP): Similarly, this clear from (MP) for $\mathsf{L}_{\mathsf{top}}$ and the fact that $u(\varphi \to \psi) = u(\varphi) \to u(\psi)$.

$\diamondsuit$

Claims 4.4.1 and 4.4.2 easily imply the following.

Claim 4.4.3. *Let $\varphi \in \mathcal{L}$. Then $\vdash_{\mathsf{L}_{\mathsf{S4A}}} t(\varphi)$ implies $\vdash_{\mathsf{L}_{\mathsf{top}}} \varphi$.*

Proof of claim. Suppose $\vdash_{\mathsf{L}_{\mathsf{S4A}}} t(\varphi)$. By Claim 4.4.2, $\vdash_{\mathsf{L}_{\mathsf{top}}} u(t(\varphi))$. By Claim 4.4.1, $\vdash_{\mathsf{L}_{\mathsf{top}}} \varphi \leftrightarrow u(t(\varphi))$. By (MP), $\vdash_{\mathsf{L}_{\mathsf{top}}} \varphi$. $\diamondsuit$

We can now show strong completeness. Suppose $\Gamma \subseteq \mathcal{L}$, $\varphi \in \mathcal{L}$ and $\Gamma \models_{\mathbb{M}_{\mathsf{top}}} \varphi$. We claim $t(\Gamma) \models_{\mathbb{M}^*_{\mathsf{S4}}} t(\varphi)$. Indeed, if $M^* \in \mathbb{M}^*_{\mathsf{S4}}$ and x is a state in M^* with $M^*, x \models t(\psi)$ for all $\psi \in \Gamma$, then with f as in Theorem 4.3.3 we have $f^{-1}(M^*), x \models \psi$ for all $\psi \in \Gamma$. Since $f^{-1}(M^*) \in \mathbb{M}_{\mathsf{int}} \cap \mathbb{M}_{\mathsf{unions}} \subseteq \mathbb{M}_{\mathsf{top}}$, $\Gamma \models_{\mathbb{M}_{\mathsf{top}}} \varphi$ gives $f^{-1}(M^*), x \models \varphi$, and thus $M^*, x \models t(\varphi)$.

By (strong) completeness of $\mathsf{L}_{\mathsf{S4A}}$ for $\mathbb{M}^*_{\mathsf{S4}}$, we get $t(\Gamma) \vdash_{\mathsf{L}_{\mathsf{S4A}}} t(\varphi)$. That is, there are $\psi_0, \dots, \psi_n \in \Gamma$ such that $\vdash_{\mathsf{L}_{\mathsf{S4A}}} t(\psi_0) \wedge \cdots \wedge t(\psi_n) \to t(\varphi)$. Since t passes over conjunctions and implications, this means $\vdash_{\mathsf{L}_{\mathsf{S4A}}} t(\psi_0 \wedge \cdots \wedge \psi_n \to \varphi)$. By Claim 4.4.3, $\vdash_{\mathsf{L}_{\mathsf{top}}} \psi_0 \wedge \cdots \wedge \psi_n \to \varphi$. Hence $\Gamma \vdash_{\mathsf{L}_{\mathsf{top}}} \varphi$, and we are done. $\square$

Just as the connection between S4 and $\mathbb{M}_{\mathsf{int}} \cap \mathbb{M}_{\mathsf{unions}}$ allowed us to obtain a complete axiomatisation of $\mathbb{M}_{\mathsf{top}}$, we can axiomatise $\mathbb{M}_{\mathsf{int}} \cap \mathbb{M}_{\mathsf{compl}}$ by considering S5. Let $\mathsf{L}_{\mathsf{int-compl}}$ be the extension of $\mathsf{L}_{\mathsf{top}}$ with the **5** axiom for $\hat{\mathsf{S}}$, which we present in the "diamond" form:

$$\mathsf{S}\neg\mathsf{S}\varphi \to \neg\mathsf{S}\varphi \quad (5_{\mathsf{S}})$$

Theorem 4.4.5. $\mathsf{L}_{\mathsf{int-compl}}$ *is sound and strongly complete with respect to* $\mathbb{M}_{\mathsf{int}} \cap \mathbb{M}_{\mathsf{compl}}$.

Proof. For soundness, we need to check that (5_{S}) is valid on $\mathbb{M}_{\mathsf{int}} \cap \mathbb{M}_{\mathsf{compl}}$. Let $M = (X, P, V)$ be closed under intersections and complements, and suppose $M, x \models \mathsf{S}\neg\mathsf{S}\varphi$. Note that $\|\mathsf{S}\varphi\|_M = \bigcap\{A \in P \mid \|\varphi\|_M \subseteq A\}$ is an intersection from P, so $\|\mathsf{S}\varphi\|_M \in P$. By closure under complements, $\|\neg\mathsf{S}\varphi\|_M \in P$ too. Hence $M, x \models \mathsf{S}\neg\mathsf{S}\varphi \wedge \mathsf{E}\neg\mathsf{S}\varphi$. By Proposition 4.1.1 (4), we get $M, x \models \neg\mathsf{S}\varphi$.

The completeness proof goes in exactly the same way as Theorem 4.4.4. Letting $\mathsf{L}_{\mathsf{S5A}}$ be the extension of $\mathsf{L}_{\mathsf{S4A}}$ with the (5_{K}) axiom $\neg\mathsf{K}\varphi \to \mathsf{K}\neg\mathsf{K}\varphi$, it can be shown that $\mathsf{L}_{\mathsf{S5A}}$ is strongly complete with respect to $\mathbb{M}^*_{\mathsf{S5}}$. With u as in the proof of Theorem 4.4.4, we have that $\vdash_{\mathsf{L}_{\mathsf{S5A}}} \varphi$ implies $\vdash_{\mathsf{L}_{\mathsf{int-compl}}} u(\varphi)$, for $\varphi \in \mathcal{L}_{\mathsf{KA}}$ (the only new part to check there is that $u(\neg\mathsf{K}\varphi \to \mathsf{K}\neg\mathsf{K}\varphi)$ is a theorem of $\mathsf{L}_{\mathsf{int-compl}}$, but this follows from (5_{S})). The remainder of the proof goes through as before, this time appealing to the bijection $g : \mathbb{M}_{\mathsf{int}} \cap \mathbb{M}_{\mathsf{compl}} \to \mathbb{M}^*_{\mathsf{S5}}$. $\square$

4.5 The Multi-source Case

So far we have been able to model the expertise of only a single source. In this section we generalise the setting to handle *multiple* sources. This allows us to consider not only the expertise of different sources individually, but also notions of *collective expertise*. For example, how may sources *combine* their expertise? Is there a suitable notion of *common expertise*? To answer these questions we take inspiration from the well-studied notions of *distributed knowledge* and *common knowledge* from epistemic logic (Fagin et al. 2003), and establish connections between collective expertise and collective knowledge.

4.5.1 Collective Knowledge

Let $\mathcal{J}$ be a finite, non-empty set of sources. Turning briefly to epistemic logic interpreted under relational semantics, we recount several notions of collective knowledge. First, a *multi-source relational model* is a triple $M^* = (X, \{R_j\}_{j\in\mathcal{J}}, V)$, where R_j is a binary relation on X for each j. Consider the following knowledge operators (Fagin et al. 2003):

- $\mathsf{K}_j\varphi$ (individual knowledge): for $j \in J$ and a formula φ, set

$$M^*, x \models \mathsf{K}_j\varphi \iff \forall y \in X : xR_jy \implies M^*, y \models \varphi.$$

  This is the straightforward adaptation of knowledge in the single-source case to the multi-source setting.

- $\mathsf{K}^{\mathsf{dist}}_J\varphi$ (distributed knowledge): for $J \subseteq \mathcal{J}$ non-empty, set

$$M^*, x \models \mathsf{K}^{\mathsf{dist}}_J\varphi \iff \forall y \in X : (x, y) \in \bigcap_{j\in J} R_j \implies M^*, y \models \varphi.$$

  That is, knowledge of φ is distributed among the sources in J if, by combining their accessibility relations R_j, all states possible at x satisfy

φ. Here the R_j are combined by taking their intersection: a state y is possible according to the group at x iff *every* source in J considers y possible at x.

- $\mathsf{K}_J^{\mathsf{sh}}\varphi$ (shared knowledge):[11] for $J \subseteq \mathcal{J}$ non-empty, set

$$M^*, x \models \mathsf{K}_J^{\mathsf{sh}}\varphi \iff \forall j \in J : M^*, x \models \mathsf{K}_j\varphi.$$

That is, a group J have shared knowledge of φ exactly when each agent in J knows φ. Thus we have $\mathsf{K}_J^{\mathsf{sh}}\varphi \equiv \bigwedge_{j\in J} \mathsf{K}_j\varphi$.

- $\mathsf{K}_J^{\mathsf{com}}\varphi$ (common knowledge): write $\mathsf{K}_J^1\varphi$ for $\mathsf{K}_J^{\mathsf{sh}}\varphi$, and for $n \in \mathbb{N}$ write $\mathsf{K}_J^{n+1}\varphi$ for $\mathsf{K}_J^{\mathsf{sh}}\mathsf{K}_J^n\varphi$. Then

$$M^*, x \models \mathsf{K}_J^{\mathsf{com}}\varphi \iff \forall n \in \mathbb{N} : M^*, x \models \mathsf{K}_J^n\varphi.$$

Here $\mathsf{K}_J^1\varphi$ says that everyone in J knows φ, $\mathsf{K}_J^2\varphi$ says that everybody in J knows that everybody in J knows φ, and so on. There is common knowledge of φ among J if this nesting of "everybody knows" holds for any order n.

In what follows we write $\mathcal{L}_{\mathsf{KA}}^{\mathcal{J}}$ for the language formed from Prop with knowledge operators K_j, $\mathsf{K}_J^{\mathsf{dist}}$, $\mathsf{K}_J^{\mathsf{sh}}$ and $\mathsf{K}_J^{\mathsf{com}}$, for $j \in \mathcal{J}$ and $J \subseteq \mathcal{J}$ non-empty, and the universal modality A.

4.5.2 Collective Expertise

Returning to expertise semantics, define a *multi-source expertise model* as a triple $M = (X, \{P_j\}_{j\in\mathcal{J}}, V)$, where $P_j \subseteq 2^X$ is the collection of expertise sets for source j. Say M is closed under intersections, unions, complements etc. if each P_j is. Since the connection between expertise and S4 knowledge (Theorem 4.3.3) holds for expertise models closed under unions and intersections, we restrict attention to this class of (multi-source) models in this section.

The counterpart of individual knowledge – individual expertise – is straightforward: we may simply introduce expertise and soundness operators E_j and S_j for each source $j \in \mathcal{J}$, and interpret $\mathsf{E}_j\varphi$ and $\mathsf{S}_j\varphi$ as in the single-source case using P_j. For notions of collective expertise and soundness, we define new collections P_J by combining the P_j in an appropriate way.

Distributed Expertise. For distributed expertise, the intuition is clear: the sources in a group J should combine their expertise collections P_j to form a larger collection P_J^{dist}. A first candidate for P_J^{dist} would therefore be $\bigcup_{j\in J} P_j$. However, since we assume each P_j is closed under unions and intersections, we suppose that each source j has the cognitive or computational capacity to

[11] In Fagin et al. (2003), shared knowledge is denoted $\mathsf{E}_J\varphi$ for "everybody knows φ". We opt to use the term "shared" knowledge to avoid conflict with our notation for expertise.

combine expertise sets $A \in P_j$ by taking unions or intersections. We argue that the same should be possible for the group J as a whole, and therefore let P_J^{dist} be the closure of $\bigcup_{j \in J} P_j$ under unions and intersections:

$$P_J^{\mathsf{dist}} = \bigcap \left\{ P' \supseteq \bigcup_{j \in J} P_j \mid P' \text{ is closed under unions and intersections} \right\}.$$

Note that P_J^{dist} is closed under unions and intersections, and $P_j \subseteq P_J^{\mathsf{dist}}$ for all $j \in J$ (in fact, P_J^{dist} is the smallest set with these properties). While P_J^{dist} depends on the model M, we suppress this from the notation.

P_J^{dist} also has a topological interpretation. As in Section 4.3, each P_j gives rise to an Alexandrov topology τ_j (where P_j are the closed sets) if it is closed under unions and intersections. By the aforementioned properties, τ_J^{dist} corresponds to the coarsest Alexandrov topology finer than each τ_j. On the other hand, since the join (in the lattice of topologies on X) of finitely many Alexandrov topologies is again Alexandrov (Steiner 1966, Theorems 2.4, 2.5), it follows that τ_J^{dist} is equal to the join $\bigvee_{j \in J} \tau_j$.

Now, recall from Theorem 4.3.3 that our semantics for expertise and soundness is connected to relational semantics via the mapping $P \mapsto R_P$ (Definition 4.3.2). The following result shows that P_J^{dist} corresponds to distributed knowledge under this mapping. For ease of notation, write R_J^{dist} for $R_{P_J^{\mathsf{dist}}}$ and R_j for R_{P_j}.

Proposition 4.5.1. *For any multi-source expertise model M and $J \subseteq \mathcal{J}$ non-empty,*

$$R_J^{\mathsf{dist}} = \bigcap_{j \in J} R_j.$$

Proof. "$\subseteq$": Suppose $x R_J^{\mathsf{dist}} y$. Let $j \in J$. We need to show $x R_j y$. Take any $A \in P_j$ such that $y \in A$. Then $A \in P_J^{\mathsf{dist}}$, so $x R_J^{\mathsf{dist}} y$ gives $x \in A$. Hence $x R_j y$.

"$\supseteq$": Suppose $(x, y) \in \bigcap_{j \in J} R_j$, i.e. $x R_j y$ for all $j \in J$. Set

$$P' = \{A \in P_J^{\mathsf{dist}} \mid y \in A \implies x \in A\} \subseteq P_J^{\mathsf{dist}}.$$

Then $P' \supseteq \bigcup_{j \in J} P_j$, since if $j \in J$ and $A \in P_j$ then $A \in P_J^{\mathsf{dist}}$ and $y \in A$ implies $x \in A$ by $x R_j y$. We claim P' is closed under intersections. Suppose $\{A_i\}_{i \in I} \subseteq P'$ and write $A = \bigcap_{i \in I} A_i$. Since $P' \subseteq P_J^{\mathsf{dist}}$ and P_J^{dist} is closed under intersections, $A \in P_J^{\mathsf{dist}}$. Suppose $y \in A$. Then $y \in A_i$ for each i, so $x \in A_i$ by the defining property of P'. Hence $x \in \bigcap_{i \in I} A_i = A$. This shows $A \in P'$ as desired. A similar argument shows that P' is also closed under unions.

We see from the definition of P_J^{dist} that $P_J^{\mathsf{dist}} \subseteq P'$, so in fact $P' = P_J^{\mathsf{dist}}$. It now follows that $x R_J^{\mathsf{dist}} y$: for any $A \in P_J^{\mathsf{dist}}$ with $y \in A$ we have $A \in P'$, so $x \in A$ also. $\square$

Common Expertise. Common expertise admits a straightforward definition: simply take the expertise sets in common with all P_j:

$$P_J^{\mathsf{com}} = \bigcap_{j \in J} P_j.$$

If each P_j is closed under unions and intersections, then so too is P_J^{com}.

At first this may appear *too* straightforward. The form of the definition is closer to *shared* knowledge than to common knowledge. But in fact, shared knowledge has *no* expertise counterpart which admits the type of connection established in Theorem 4.3.3. Indeed, shared knowledge may fail positive introspection (axiom **4**: $\mathsf{K}\varphi \to \mathsf{K}\mathsf{K}\varphi$), but we have seen that the knowledge derived from expertise and soundness satisfies S4 (when the collection of expertise sets is closed under unions and intersections).

However, this problem is only apparent in the translation of $\mathsf{S}\varphi$ as $\neg\mathsf{K}\neg\varphi$. For our translation of $\mathsf{E}\varphi$ as $\mathsf{A}(\neg\varphi \to \mathsf{K}\neg\varphi)$, the universal quantification via A dissolves the differences between shared and common knowledge.

Proposition 4.5.2. *Let $\varphi \in \mathcal{L}_{\mathsf{KA}}^{\mathcal{J}}$ and let $J \subseteq \mathcal{J}$ be non-empty. Then*

$$\mathsf{A}(\neg\varphi \to \mathsf{K}_J^{\mathsf{com}}\neg\varphi) \equiv \mathsf{A}(\neg\varphi \to \mathsf{K}_J^{\mathsf{sh}}\neg\varphi).$$

Proof. Let $M^* = (X, \{R_j\}_{j \in \mathcal{J}}, V)$ be a multi-source relational model. Since $\mathsf{K}_J^{\mathsf{com}}\psi \to \mathsf{K}_J^{\mathsf{sh}}\psi$ is valid for any ψ, the left-to-right implication of the above equivalence is straightforward.

For the right-to-left implication, suppose $M^*, x \models \mathsf{A}(\neg\varphi \to \mathsf{K}_J^{\mathsf{sh}}\neg\varphi)$. We show by induction that $M^*, x \models \mathsf{A}(\neg\varphi \to \mathsf{K}_J^{n}\neg\varphi)$ for all $n \in \mathbb{N}$, from which the result follows.

The base case $n = 1$ is given, since $\mathsf{K}_J^{1}\neg\varphi = \mathsf{K}_J^{\mathsf{sh}}\neg\varphi$. For the inductive step, suppose $M^*, x \models \mathsf{A}(\neg\varphi \to \mathsf{K}_J^{n}\neg\varphi)$. Take $y \in X$ such that $M^*, y \models \neg\varphi$. Let $j \in J$. Take $z \in X$ such that yR_jz. From the initial assumption we have $M^*, y \models \mathsf{K}_J^{\mathsf{sh}}\neg\varphi$, so $M^*, y \models \mathsf{K}_j\neg\varphi$ and thus $M^*, z \models \neg\varphi$. By the inductive hypothesis, $M^*, z \models \mathsf{K}_J^{n}\neg\varphi$. This shows that $M^*, y \models \mathsf{K}_j\mathsf{K}_J^{n}\neg\varphi$ for all $j \in J$, and thus $M^*, y \models K_J^{n+1}\neg\varphi$. Hence $M^*, x \models \mathsf{A}(\neg\varphi \to \mathsf{K}_J^{n+1}\neg\varphi)$ as required. $\square$

Proposition 4.5.2 shows that when interpreting collective expertise on φ as collective refutation of φ whenever φ is false, there is no difference between using common knowledge and just shared knowledge.

We now confirm that P_J^{com} does indeed correspond to common knowledge. First we recall a well-known result from Fagin et al. (2003). In what follows, write $R^+ = \bigcup_{n \in \mathbb{N}} R^n$ for the transitive closure of R.

Lemma 4.5.1 (Fagin et al. (2003), Lemma 2.2.1)**.** *Let $M^* = (X, \{R_j\}_{j \in \mathcal{J}}, V)$ be a multi-source relational model and $J \subseteq \mathcal{J}$ non-empty. Write $R' = \left(\bigcup_{j \in J} R_j\right)^+$. Then for all $x \in X$ and $\varphi \in \mathcal{L}_{\mathsf{KA}}^{\mathcal{J}}$:*

$$M^*, x \models \mathsf{K}_J^{\mathsf{com}}\varphi \iff \forall y \in X : xR'y \implies M^*, y \models \varphi.$$

By Lemma 4.5.1, common knowledge has an interpretation in terms of the usual relational semantics for knowledge, where we use the transitive closure of the union of the accessibility relations of the sources in J. Writing R_J^{com} for $R_{P_J^{\mathsf{com}}}$, we have the following.

Proposition 4.5.3. *Let M be a multi-source model closed under unions and intersections. Then for $J \subseteq \mathcal{J}$ non-empty, $R_J^{\mathsf{com}} = \left(\bigcup_{j \in J} R_j\right)^+$.*

Proof. Write $R' = (\bigcup_{j \in J} R_j)^+$. Note that R_J^{com} is reflexive and transitive by Lemma 4.3.2 (1). R' is transitive by its definition as a transitive closure, and reflexive since each R_j is (and $J \neq \emptyset$). It is therefore sufficient by Lemma 4.3.1 to show that any set is downwards closed wrt R_J^{com} iff it is downwards closed wrt R'. Since each P_j is closed under unions and intersections, so too is P_J^{com}. Using Lemma 4.3.2 (2), we have

$$\begin{aligned} A \text{ downwards closed wrt } R_J^{\mathsf{com}} &\iff A \in P_J^{\mathsf{com}} \\ &\iff \forall j \in J : A \in P_j \\ &\iff \forall j \in J : A \text{ downwards closed wrt } R_j \\ &\iff A \text{ downwards closed wrt } \bigcup_{j \in J} R_j \\ &\iff A \text{ downwards closed wrt } R' \end{aligned}$$

where the last step uses the fact that A is downwards closed with respect to some relation if and only if it is downwards closed with respect to the transitive closure. This completes the proof. □

Collective semantics. We now formally define the syntax and semantics of collective expertise. Let $\mathcal{L}^{\mathcal{J}}$ be the language defined by the following grammar:

$$\varphi ::= p \mid \varphi \wedge \varphi \mid \neg\varphi \mid \mathsf{E}_j\varphi \mid \mathsf{S}_j\varphi \mid \mathsf{E}_J^g\varphi \mid \mathsf{S}_J^g\varphi \mid \mathsf{A}\varphi$$

for $p \in \mathsf{Prop}$, $j \in \mathcal{J}$, $g \in \{\mathsf{dist}, \mathsf{com}\}$ and $J \subseteq \mathcal{J}$ non-empty. For a multi-source expertise model $M = (X, \{P_j\}_{j \in \mathcal{J}}, V)$, define the satisfaction relation as before for atomic propositions, propositional connectives and A, and set

$$\begin{array}{llll} M, x & \models \mathsf{E}_j\varphi & \iff \|\varphi\|_M \in P_j & \\ M, x & \models \mathsf{E}_J^g\varphi & \iff \|\varphi\|_M \in P_J^g & (g \in \{\mathsf{dist}, \mathsf{com}\}) \\ M, x & \models \mathsf{S}_j\varphi & \iff \forall A \in P_J : \|\varphi\|_M \subseteq A \implies x \in A & \\ M, x & \models \mathsf{S}_J^g\varphi & \iff \forall A \in P_J^g : \|\varphi\|_M \subseteq A \implies x \in A & (g \in \{\mathsf{dist}, \mathsf{com}\}) \end{array}$$

Note that expertise and soundness are interpreted as before, but with respect to different collections P. Consequently, the interactions shown in Proposition 4.1.1 still hold for individual and collective notions of expertise and soundness.

Example 4.5.1. *Extending Examples 4.1.1 and 4.1.2, consider* $\mathcal{J} = \{\mathsf{econ}, \mathsf{dr}, \mathsf{analyst}\}$, *where* econ *is the economist,* dr *is a doctor with expertise on* i *only, and* analyst *has access to aggregate data distinguishing three levels of virus activity: minimal* $(\neg i \wedge \neg d)$, *high* $((i \vee d) \wedge \neg(i \wedge d))$ *and very high* $(i \wedge d)$. *This can be modelled by a multi-source model* M *with* X, V *and* P_{econ} *as in Example 4.1.2, and* $P_{\mathsf{dr}} = \{\emptyset, X, \{ipd, ip, id, i\}, \{pd, p, d, \emptyset\}\}$, P_{analyst} *is the closure under unions of* $\{\emptyset, X, \{ipd, id\}, \{ip, pd, i, d\}, \{p, \emptyset\}\}$.

Note that neither dr *nor* analyst *have expertise on* d *individually. However, if* dr *can communicate whether or not* i *holds, this gives* analyst *enough information to disambiguate the "high activity" case and therefore determine the truth value of* d. *Indeed, we have* $\|d\| = \|i \wedge d\| \cup (\|i \vee d\| \setminus \|i \wedge d\| \cap \|\neg i\|)$, *which is formed by unions and intersections from* $P_{\mathsf{dr}} \cup P_{\mathsf{analyst}}$, *and thus* $\|d\| \in P^{\mathsf{dist}}_{\{\mathsf{dr},\mathsf{analyst}\}}$. *Hence* $M \models \mathsf{E}^{\mathsf{dist}}_{\{\mathsf{dr},\mathsf{analyst}\}} d$. *Similarly,* dr *and* analyst *have distributed expertise on* $\neg d$. *Bringing back* econ, *the grand coalition* $\mathcal{J}$ *has distributed expertise on the original report* $p \wedge \neg d$ *from Example 4.1.1. Consequently, the report is no longer sound at "actual" state* idp: *all sources together have sufficient expertise to know it is false.*

The following validities express properties specific to collective expertise.

Proposition 4.5.4. *The following formulas are valid.*

1. *For* $j \in J$, $\mathsf{E}_j \varphi \to \mathsf{E}^{\mathsf{dist}}_J \varphi$
2. $\mathsf{E}^{\mathsf{com}}_J \varphi \leftrightarrow \bigwedge_{j \in J} \mathsf{E}_j \varphi$
3. $\mathsf{S}^{\mathsf{com}}_J \varphi \leftrightarrow \bigvee_{j \in J} \mathsf{S}_j \mathsf{S}^{\mathsf{com}}_J \varphi$
4. $\mathsf{E}^{\mathsf{dist}}_{\{j\}} \varphi \leftrightarrow \mathsf{E}_j \varphi$ *is valid on* $\mathbb{M}^{\mathcal{J}}_{\mathsf{int}} \cap \mathbb{M}^{\mathcal{J}}_{\mathsf{unions}}$

Proof. We prove only (3); the others are straightforward. The right implication is valid since $\psi \to \mathsf{S}_j \psi$ is, with ψ set to $\mathsf{S}^{\mathsf{com}}_J \varphi$ and $j \in J$ arbitrary (recall J is non-empty). For the left implication, suppose there is $j \in J$ with $M, x \models \mathsf{S}_j \mathsf{S}^{\mathsf{com}}_J \varphi$. Then $x \in \bigcap \{A \in P_j \mid \|\mathsf{S}^{\mathsf{com}}_J \varphi\|_M \subseteq A\}$. Now take $B \in P^{\mathsf{com}}_J$ such that $\|\varphi\|_M \subseteq B$. Note that if $y \in \|\mathsf{S}^{\mathsf{com}}_J \varphi\|$ then $y \in B$ by the definition of the semantics for $\mathsf{S}^{\mathsf{com}}_J$, so $\|\mathsf{S}^{\mathsf{com}}_J \varphi\|_M \subseteq B$. Since $B \in P^{\mathsf{com}}_J \subseteq P_j$, we get $x \in B$. This shows $M, x \models \mathsf{S}^{\mathsf{com}}_J \varphi$. □

Validity (3) comes from the *fixed-point axiom* for common knowledge: $\mathsf{K}^{\mathsf{com}}_J \varphi \leftrightarrow \mathsf{K}^{\mathsf{sh}}_J (\varphi \wedge \mathsf{K}^{\mathsf{com}}_J \varphi)$. Our version says $\mathsf{S}^{\mathsf{com}}_J \varphi$ is a fixed-point of the function $\theta \mapsto \bigvee_{j \in J} \mathsf{S}_j \theta$. In words, φ is true up to lack of *common* expertise iff there is some source for whom $\mathsf{S}^{\mathsf{com}}_J \varphi$ is true up to their lack of (individual) expertise.

As promised, there is a tight link between our notions of collective expertise and knowledge. Define a translation $t : \mathcal{L}^{\mathcal{J}} \to \mathcal{L}^{\mathcal{J}}_{\mathsf{KA}}$ inductively by

$$\begin{array}{lll} t(\mathsf{E}_j\varphi) & = \mathsf{A}(\neg t(\varphi) \to \mathsf{K}_j\neg t(\varphi)) & \\ t(\mathsf{E}^g_J\varphi) & = \mathsf{A}(\neg t(\varphi) \to \mathsf{K}^g_J\neg t(\varphi)) & (g \in \{\mathsf{dist}, \mathsf{com}\}) \\ t(\mathsf{S}_j\varphi) & = \neg\mathsf{K}_j\neg t(\varphi) & \\ t(\mathsf{S}^g_J\varphi) & = \neg\mathsf{K}^g_J\neg t(\varphi) & (g \in \{\mathsf{dist}, \mathsf{com}\}) \end{array}$$

where the other cases are as for t in Section 4.3. This is essentially the same translation as before, but with the various types of expertise and soundness matched with their knowledge counterparts. We have an analogue of Theorem 4.3.3.

Theorem 4.5.1. *The mapping $f : \mathbb{M}^{\mathcal{J}}_{\mathsf{int}} \cap \mathbb{M}^{\mathcal{J}}_{\mathsf{unions}} \to \mathbb{M}^{\mathcal{J}}_{\mathsf{S4}}$ given by $(X, \{P_j\}_{j\in\mathcal{J}}, V) \mapsto (X, \{R_{P_j}\}_{j\in\mathcal{J}}, V)$ is bijective, and for $x \in X$ and $\varphi \in \mathcal{L}^{\mathcal{J}}$:*

$$M, x \models \varphi \iff f(M), x \models t(\varphi).$$

Moreover, the restriction of this map to $\mathbb{M}^{\mathcal{J}}_{\mathsf{int}} \cap \mathbb{M}^{\mathcal{J}}_{\mathsf{compl}}$ is a bijection into $\mathbb{M}^{\mathcal{J}}_{\mathsf{S5}}$.

Proof. That the map is bijective follows easily from Theorems 4.3.1 and 4.3.2. For the stated property we proceed by induction on $\mathcal{L}^{\mathcal{J}}$ formulas. As in Theorem 4.3.3, the cases for atomic propositions, propositional connectives and A are straightforward. For expertise and soundness, the argument in the proof of Theorem 4.3.3 showed that $\mathsf{E}\varphi$ and $\mathsf{S}\varphi$ interpreted via some collection P is equivalent to $t(\mathsf{E}\varphi)$ and $t(\mathsf{S}\varphi)$ interpreted wrt relational semantics via R_P. It is therefore sufficient to show that for each notion of individual and collective expertise interpreted in M via P, its corresponding notion of individual or collective knowledge (used in the translation t) is interpreted in $f(M)$ via R_P. This is self-evident for individual expertise. For distributive expertise this was shown in Proposition 4.5.1. For common expertise this was shown in Lemma 4.5.1 and Proposition 4.5.3. □

Theorem 4.5.1 can be used to adapt any sound and complete axiomatisation for $\mathbb{M}^{\mathcal{J}}_{\mathsf{S4}}$ (resp., $\mathbb{M}^{\mathcal{J}}_{\mathsf{S5}}$) over the language $\mathcal{L}^{\mathcal{J}}_{\mathsf{KA}}$ to obtain an axiomatisation for $\mathbb{M}^{\mathcal{J}}_{\mathsf{int}} \cap \mathbb{M}^{\mathcal{J}}_{\mathsf{unions}}$ (resp., $\mathbb{M}^{\mathcal{J}}_{\mathsf{int}} \cap \mathbb{M}^{\mathcal{J}}_{\mathsf{compl}}$) over $\mathcal{L}^{\mathcal{J}}$, in the same way as we did earlier when adapting S4 and S5 in Theorems 4.4.4 and 4.4.5.

4.6 Dynamic Extension

So far our picture has been entirely static. We cannot speak of expertise changing over time, nor of the information in a model changing via announcements from sources. To remedy this, we extend the framework with two *dynamic* operators: one to account for *increases in expertise* – e.g. after a process of learning or acquisition of new evidence – and one to model *sound announcements*. For simplicity, we return to the single-source case.

4.6.1 Expertise Increase

As a source interacts with the world over time, they may learn to make more distinctions between possible states of the world, and thereby increase their expertise. Leaving the particulars of the learning mechanism unspecified, we study only the end result: the source's expertise collection P is expanded to include a new set A.

However, this may not be so simple as setting $P' = P \cup \{A\}$ in light of the closure properties that may be imposed P. As remarked in Section 4.2, closure conditions correspond to assumptions about the source's cognitive or computational capabilities. It seems natural that if the source has the ability to combine sets in P by taking intersections, for example, then they should also be able to do after the learning, i.e. P' should also be closed under intersections. Thus, the new collection P' should inherit any closure properties from P, while extending $P \cup \{A\}$. In principle, we could therefore consider an expertise increase operation for *each* combination of closure properties.

For concreteness we will not do this, and will instead focus on the class $\mathbb{M}_{\mathsf{int}}$ of models closed under intersections. Conceptually, this is a minimal requirement, since we argued in section Section 4.2 that closure under intersections is a natural property. There are also technical advantages: we will later show that closure under intersections allows us to find reduction axioms which allow the formulas involving expertise increase to be equivalently expressed in the static language.

Definition 4.6.1. *Given an expertise model $M = (X, P, V)$ and a formula φ, define the model $M^{+\varphi} = (X, P^{+\varphi}, V)$ by setting*

$$P^{+\varphi} = \left\{ \bigcap \mathcal{A} \mid \mathcal{A} \subseteq P \cup \{\|\varphi\|_M\} \right\}.$$

That is, $P^{+\varphi}$ is obtained by adding $\|\varphi\|_M$ to P and closing under intersections.

Syntactically, we introduce formulas of the form $[+\varphi]\psi$, which are to be read as "ψ holds after the source gains expertise on φ". The truth condition for $[+\varphi]\psi$ in a model M is defined in terms of $M^{+\varphi}$:

$$M, x \models [+\varphi]\psi \iff M^{+\varphi}, x \models \psi.$$

If $\mathcal{L}_0$ denotes the propositional language built from Prop, then $[+\alpha]\mathsf{E}\alpha$ is valid for all $\alpha \in \mathcal{L}_0$. That is, expertise increase is successful for any propositional formula. However, this is not the case for general formulas $\varphi \in \mathcal{L}$. This comes from the fact that expertise is represented *semantically* via sets of states. The operator $[+\varphi]$ represents the source obtaining expertise on the set of φ states, where φ is interpreted *before the increase took place.* If φ refers to expertise (with E or S) then the meaning of φ may change after the increase. For example, consider the model $M = (X, P, V)$ with

$$\begin{aligned} X &= \{1, 2, 3, 4\} \\ P &= \{\emptyset, X, \{1, 3\}\} \\ V(p) &= \{1\} \\ V(q) &= \{2, 3\} \end{aligned}$$

Then, with $\varphi = p \vee (q \wedge \neg \mathsf{S}p)$ we have $M, 1 \not\models [+\varphi]\mathsf{E}\varphi$.[12] This counterexample is reminiscent of *Moore sentences* as formalised in Dynamic Epistemic Logic; e.g. an agent cannot know $p \wedge \neg \mathsf{K}p$ ("p is true but I do not know it") after this is truthfully announced (Baltag and Smets 2008).

Next we give reduction axioms to express any formula involving $[+\varphi]$ by an equivalent formula in the static language $\mathcal{L}$.

Proposition 4.6.1. *The following formulas are valid on* $\mathbb{M}$*:*

$$\begin{array}{rcl} [+\varphi]p & \leftrightarrow & [+\varphi]p \\ [+\varphi](\psi \wedge \theta) & \leftrightarrow & [+\varphi]\psi \wedge [+\varphi]\theta \\ [+\varphi]\neg\psi & \leftrightarrow & \neg[+\varphi]\psi \\ [+\varphi]\mathsf{A}\psi & \leftrightarrow & \mathsf{A}[+\varphi]\psi \\ [+\varphi]\mathsf{S}\psi & \leftrightarrow & \mathsf{S}[+\varphi]\psi \wedge (\mathsf{A}([+\varphi]\psi \to \varphi) \to \varphi) \\ [+\varphi]\mathsf{E}\psi & \leftrightarrow & \mathsf{A}((\mathsf{S}[+\varphi]\psi \wedge (\mathsf{A}([+\varphi]\psi \to \varphi) \to \varphi)) \to [+\varphi]\psi) \end{array}$$

Proof. The cases for atomic propositions, propositional connectives and A are straightforward. We show the reduction axiom for S. Let $M = (X, P, V)$ be a model and $x \in X$.

"$\to$": Suppose $M, x \models [+\varphi]\mathsf{S}\psi$. Then $M^{+\varphi}, x \models \mathsf{S}\psi$. Hence

$$x \in \bigcap\{A \in P^{+\varphi} \mid \|\psi\|_{M^{+\varphi}} \subseteq A\} \quad (*)$$

Note that $\|\psi\|_{M^{+\varphi}} = \|[+\varphi]\psi\|_M$. Now take $A \in P$ such that $\|[+\varphi]\psi\|_M \subseteq A$. Since $P \subseteq P^{+\varphi}$, we get $x \in A$ from $(*)$. Hence $M, x \models \mathsf{S}[+\varphi]\psi$.

Now suppose $M, x \models \mathsf{A}([+\varphi]\psi \to \varphi)$. Then $\|[+\varphi]\psi\|_M \subseteq \|\varphi\|_M$, so $\|\psi\|_{M^{+\varphi}} \subseteq \|\varphi\|_M$. Since $\|\varphi\|_M \in P^{+\varphi}$, we get $x \in \|\varphi\|_M$ from $(*)$, i.e. $M, x \models \varphi$ as required.

"$\leftarrow$": Suppose $M, x \models \mathsf{S}[+\varphi]\psi$ and $M, x \models \mathsf{A}([+\varphi]\psi \to \varphi) \to \varphi$. Take $A \in P^{+\varphi}$ such that $\|\psi\|_{M^{+\varphi}} \subseteq A$. Then $\|[+\varphi]\psi\|_M \subseteq A$. By definition of $P^{+\varphi}$, there is a collection $\mathcal{A} \subseteq P \cup \{\|\varphi\|_M\}$ such that $A = \bigcap \mathcal{A}$. Let $B \in \mathcal{A}$. If $B \in P$, then $\|[+\varphi]\psi\|_M \subseteq A \subseteq B$ and $M, x \models \mathsf{S}[+\varphi]\psi$ give $x \in B$. Otherwise, $B = \|\varphi\|_M$. Hence $\|[+\varphi]\psi\|_M \subseteq \|\varphi\|_M$, so $M, x \models \mathsf{A}([+\varphi]\psi \to \varphi)$. By the second assumption, we get $M, x \models \varphi$, i.e. $x \in \|\varphi\|_M = B$. We have now shown that $x \in \bigcap \mathcal{A} = A$, and thus $M^{+\varphi}, x \models \mathsf{S}\psi$ and $M, x \models [+\varphi]\mathsf{S}\psi$.

[12] In detail, we have $\|\varphi\|_M = \{1, 2\}$, so $P^{+\varphi} = \{\emptyset, X, \{1, 3\}, \{1, 2\}, \{1\}\}$. Then $\|\varphi\|_{M^{+\varphi}} = \{1, 2, 3\} \notin P^{+\varphi}$, so $M^{+\varphi}, 1 \not\models \mathsf{E}\varphi$.

For the reduction axiom for E, note that since $M^{+\varphi} \in \mathbb{M}_{\mathsf{int}}$ we have $M^{+\varphi}, x \models E\psi$ iff $M^{+\varphi}, x \models \mathsf{A}(\mathsf{S}\psi \to \psi)$. Using the reduction axioms for A and S (and the reduction axiom for the implication, derived from those for $\neg$ and $\wedge$), we obtain the desired equivalence. $\square$

Note that only the reduction axiom for $[+\varphi]\mathsf{E}\psi$ requires $M^{+\varphi}$ to be closed under intersections.

4.6.2 Sound Announcements

In logics of public announcement (Plaza 2007; van Ditmarsch, van der Hoek, and Kooi 2008), the dynamic operator $[!\varphi]$ represents a *public* and *truthful* announcement of φ; the formula $[!\varphi]\psi$ is read as "after φ is announced, ψ holds". Such an announcement changes the information available in a model: after the announcement, all $\neg\varphi$ states are eliminated.

Since the premise of our work is to deal with non-expert sources, the truthfulness requirement is too strong for an announcement operator in our setting. Instead, we consider *sound announcements*: the source may announce φ whenever φ is sound at the current state. That is, the source may announce any (possibly false) statement which is true up to their lack of expertise.

Such an announcement is denoted syntactically by $[?\varphi]$. As with the expertise increase operator, we define a model update operation $M \mapsto M^{?\varphi}$.

It is clear how one should define new set of states: since the announcement tells us φ is sound, we eliminate unsound states by setting $X^{?\varphi} = \|\mathsf{S}\varphi\|_M$. The valuation is also straightforward, since announcements should not change the meaning of atomic propositions.

What about the new expertise collection $P^{?\varphi}$? If we restrict attention to models closed under intersections, as we did for expertise increase, then a natural choice is to simply restrict each $A \in P$ to $X^{?\varphi}$ by intersection. Since $X^{?\varphi} = \|\mathsf{S}\varphi\|_M = \bigcap\{B \in P \mid \|\varphi\|_M \subseteq B\}$, by the closure property we will have $P^{?\varphi} \subseteq P$, so that announcements do not increase expertise. This assumption will also permit us to find reduction axioms later on.

Definition 4.6.2. *Let $M = (X, P, V)$ be an expertise model. For a formula φ, define the model $M^{?\varphi} = (X^{?\varphi}, P^{?\varphi}, V^{?\varphi})$ by setting*

$$
\begin{aligned}
X^{?\varphi} &= \|\mathsf{S}\varphi\|_M \\
P^{?\varphi} &= \{A \cap X^{?\varphi} \mid A \in P\} \\
V^{?\varphi} &= V(p) \cap X^{?\varphi}
\end{aligned}
$$

Semantically, the truth condition for $[?\varphi]\psi$ is as follows.

$$M, x \models [?\varphi]\psi \iff (M, x \models \mathsf{S}\varphi \implies M^{?\varphi}, x \models \psi).$$

Here we have the precondition that $\mathsf{S}\varphi$ is true: if φ is unsound, $[?\varphi]\psi$ is true for *any* ψ. Note that a sound announcement of φ can also be seen as a public (*truthful*) announcement of $\mathsf{S}\varphi$.

Example 4.6.1. *The report of the economist in Example 4.1.1 can be modelled as* $[?(p \wedge \neg d)]$*. Note that, with* M *as in Example 4.1.2,* $\|\mathsf{S}(p \wedge \neg d)\|_M = \|p\|_M$*. The updated model* $M^{?(p \wedge \neg d)}$ *therefore consists only of the bottom half of* M *as shown in Fig. 4.1. We see that* $M, idp \models [?(p \wedge \neg d)]d$ *– showing that even propositional announcements can "fail" due to lack of expertise – and* $M \models [?(p \wedge \neg d)]\mathsf{A}p$ *– showing that the parts of the report on which the source does have expertise are* always *true after their announcement.*

As with the expertise increase operator, sound announcements remain sound for purely propositional formulas $\alpha \in \mathcal{L}_0$: $[?\alpha]\mathsf{S}\alpha$ is valid on $\mathbb{M}_{\mathsf{int}}$. This is not true for general formulas $\varphi \in \mathcal{L}$ which may refer to expertise itself. For example, in the model $M = (X, P, V) \in \mathbb{M}_{\mathsf{int}}$ given by $X = \{1, 2, 3, 4\}$, $P = \{\emptyset, X, \{1\}, \{2\}, \{1, 2, 3\}\}$, $V(p) = \{1, 2\}$ and $V(q) = \{2, 4\}$, with $\varphi = p \wedge \neg \mathsf{E}q$ we have $M, 1 \not\models [?\varphi]\mathsf{S}\varphi$.

The following reduction axioms allow formulas involving announcements to be expressed in the static language.

Proposition 4.6.2. *The following formulas are valid on* $\mathbb{M}$*:*

$$\begin{aligned}
[?\varphi]p \quad &\leftrightarrow \quad \mathsf{S}\varphi \rightarrow p \\
[?\varphi](\psi \wedge \theta) \quad &\leftrightarrow \quad [?\varphi]\psi \wedge [?\varphi]\theta \\
[?\varphi]\neg\psi \quad &\leftrightarrow \quad \mathsf{S}\varphi \rightarrow \neg[?\varphi]\psi \\
[?\varphi]\mathsf{A}\psi \quad &\leftrightarrow \quad \mathsf{S}\varphi \rightarrow \mathsf{A}[?\varphi]\psi \\
[?\varphi]\mathsf{S}\psi \quad &\leftrightarrow \quad \mathsf{S}\varphi \rightarrow \mathsf{S}(\mathsf{S}\varphi \wedge [?\varphi]\psi)
\end{aligned}$$

and the following is valid on $\mathbb{M}_{\mathsf{int}}$*:*

$$[?\varphi]\mathsf{E}\psi \quad \leftrightarrow \quad \mathsf{S}\varphi \rightarrow \mathsf{E}(\mathsf{S}\varphi \wedge [?\varphi]\psi)$$

Proof. The cases of atomic propositions, propositional connectives and the universal modality A are straightforward.

For the reduction axiom for S, first note that $\|\psi\|_{M^{?\varphi}} = \|\mathsf{S}\varphi \wedge [?\varphi]\psi\|_M$. We need to show that $M, x \models [?\varphi]\mathsf{S}\psi$ iff $M, x \models \mathsf{S}\varphi \rightarrow \mathsf{S}(\mathsf{S}\varphi \wedge [?\varphi]\psi)$. If $M, x \not\models \mathsf{S}\varphi$ this is clear. Otherwise $x \in \|\mathsf{S}\varphi\|_M$, and we have

$$\begin{aligned}
M, x \models [?\varphi]\mathsf{S}\psi &\iff M^{?\varphi}, x \models \mathsf{S}\psi \\
&\iff \forall B \in P^{?\varphi} : \|\psi\|_{M^{?\varphi}} \subseteq B \implies x \in B \\
&\iff \forall A \in P : \|\mathsf{S}\varphi \wedge [?\varphi]\psi\|_M \subseteq A \cap \|\mathsf{S}\varphi\|_M \implies x \in A \cap \|\mathsf{S}\varphi\|_M \\
&\iff \forall A \in P : \|\mathsf{S}\varphi \wedge [?\varphi]\psi\|_M \subseteq A \implies x \in A \\
&\iff M, x \models \mathsf{S}(\mathsf{S}\varphi \wedge [?\varphi]\psi)
\end{aligned}$$

and the result follows.

For the E reduction axiom, take $M \in \mathbb{M}_{\mathsf{int}}$. Again, suppose without loss of generality that $x \in \|\mathsf{S}\varphi\|_M$. Then we have

$$\begin{aligned} M, x \models [?\varphi]\mathsf{E}\psi &\iff M^{?\varphi}, x \models \mathsf{E}\psi \\ &\iff \|\psi\|_{M^{?\varphi}} \in P^{?\varphi} \\ &\iff \|\mathsf{S}\varphi \wedge [?\varphi]\psi\|_M \in P^{?\varphi} \\ &\iff \|\mathsf{S}\varphi \wedge [?\varphi]\psi\|_M \in P \\ &\iff M, x \models \mathsf{E}(\mathsf{S}\varphi \wedge [?\varphi]\psi) \end{aligned}$$

where the forwards direction of the penultimate equivalence holds since $P^{?\varphi} \subseteq P$ when M is closed under intersections, and the backwards direction holds since $\|\mathsf{S}\varphi \wedge [?\varphi]\psi\|_M \subseteq \|\mathsf{S}\varphi\|_M = X^{?\varphi}$. It follows that $M, x \models [?\varphi]\mathsf{E}\psi$ iff $M, x \models \mathsf{S}\varphi \to \mathsf{E}(\mathsf{S}\varphi \wedge [?\varphi]\psi)$, as required. $\square$

To conclude, we note some interesting validities involving the dynamic operators and their interaction.

Proposition 4.6.3. *For any $\alpha, \beta \in \mathcal{L}_0$, the following formulas are valid on $\mathbb{M}_{\mathsf{int}}$:*

1. $\mathsf{E}\alpha \leftrightarrow \mathsf{A}[?\alpha]\alpha$
2. $\mathsf{A}(\alpha \to \beta) \to [+\beta][?\alpha]\beta$
3. $[+\alpha][?\alpha]\alpha$

Proof.

1. Using the reduction axioms for atomic propositions, conjunctions and negations, one can show by induction that $[?\varphi]\alpha$ is equivalent to $\mathsf{S}\varphi \to \alpha$. Applying this with $\varphi = \alpha$, we have that $\mathsf{A}[?\alpha]\alpha$ is equivalent to $\mathsf{A}(\mathsf{S}\alpha \to \alpha)$, which is equivalent to $\mathsf{E}\alpha$ for models closed under intersections.

2. We use the following fact, whose proof is straightforward by induction on $\mathcal{L}_0$ formulas.

   - For $\alpha \in \mathcal{L}_0$, $\varphi \in \mathcal{L}$ and any model M, $\|\alpha\|_{M^{+\varphi}} = \|\alpha\|_M$ and $\|\alpha\|_{M^{?\varphi}} = \|\alpha \wedge \mathsf{S}\varphi\|_M$.

   Now, take $M = (X, P, V) \in \mathbb{M}_{\mathsf{int}}$, $x \in X$, and suppose $M, x \models \mathsf{A}(\alpha \to \beta)$. Then $\|\alpha\|_M \subseteq \|\beta\|_M$.

   We need to show $M, x \models [+\beta][?\alpha]\beta$, i.e. $M^{+\beta}, x \models [?\alpha]\beta$. Suppose $M^{+\beta}, x \models \mathsf{S}\alpha$. To show $(M^{+\beta})^{?\alpha}, x \models \beta$, we need

   $$x \in \|\beta\|_{(M^{+\beta})^{?\alpha}} = \|\beta \wedge \mathsf{S}\alpha\|_{M^{+\beta}}$$

   where the equality follows from the claim above. By assumption $M^{+\beta}, x \models \mathsf{S}\alpha$, so we only need to show $M^{+\beta}, x \models \beta$.

Since $[+\beta]\mathsf{E}\beta$ is valid in M, we have $M^{+\beta}, x \models \mathsf{E}\beta$. From Proposition 4.1.1 (3), $M^{+\beta}, x \models \mathsf{A}(\alpha \to \beta) \to (\mathsf{S}\alpha \wedge \mathsf{E}\beta \to \beta)$. But from the above claim and $\|\alpha\|_M \subseteq \|\beta\|_M$ we have $\|\alpha\|_{M^{+\beta}} \subseteq \|\beta\|_{M^{+\beta}}$, i.e. $M^{+\beta}, x \models \mathsf{A}(\alpha \to \beta)$. Hence $M^{+\beta}, x \models \beta$, and we are done.

3. Taking $\beta = \alpha$, this validity follows from (2).

□

In words, (1) says that expertise on a propositional formula α is equivalent to the guarantee that α is true whenever it is soundly announced. (2) is essentially a reformulation of Proposition 4.1.1 (3); it says that if β is logically weaker than α, gaining expertise on β ensures that β is at least true after a sound announcement of the stronger formula α. (3) is the special case of (2) with $\beta = \alpha$, which says that α is true following a sound announcement after the sources gains expertise on α.

4.7 Conclusion

Summary. This chapter presented a modal logic framework to reason about the expertise of information sources and soundness of information. We investigated both conceptual and technical issues, establishing completeness results for various classes of expertise models. The connection with epistemic logic showed how expertise and soundness may be given precise interpretations in terms of knowledge; if expertise is closed under intersections and unions this results in S4 knowledge, and closure under complements strengthens this to S5. The framework was then extended to handle multiple sources, permitting the study of several notions of collective expertise. Finally, we considered dynamic operators to model evolving expertise and sound announcements.

Limitations and future work. On a technical level, some open questions remain. For example, can frames closed under arbitrary unions be be expressed in our language, as other closure properties were expressed in Proposition 4.2.1? Similarly, can one axiomatise the class of models closed under arbitrary unions, without also requiring closure under intersections? One could also consider computational properties, such as decidability and the complexity of the satisfiability problem.

There are also conceptual limitations and areas for future study. Firstly, our notion of expertise is absolute: either the source is an expert on φ or they are not. In reality things are more nuanced, and source may have varying levels of expertise. Our assumption that expertise is independent of the actual state of the world could also be considered too strong, since it forbids any possibility of *context-dependent* expertise. As a somewhat contrived example, the economist in our running examples may have expertise on p in ordinary times, but not if they are suffering from the virus which affects cognitive ability.

Outlook. Equipped with the notions of expertise and soundness from this chapter, the following chapter poses a belief change problem – in the style of AGM revision (Alchourrón, Gärdenfors, and Makinson 1985) and belief merging (Konieczny and Pino Pérez 2002) – in which expertise is not assumed to be known upfront, but must be estimated from a sequence of reports. For simplicity we dispense with some of the generality of this chapter, by (i) considering only finite models whose states are the propositional valuations over a fixed, finite set of variables; and (ii) assuming expertise collections are closed under both intersections and complements. By Theorem 4.3.2, such models are in one-to-one correspondence with S5 relational models, so that their corresponding binary relation R_P is an equivalence relation. Equivalently, each collection P closed under intersections and complements corresponds to a partition Π_P over the set of states, whose cells are simply the equivalence classes of R_P. Since one can express the semantic conditions for expertise and soundness directly in terms of this partition, in what follows we in fact take the partition Π_P as primitive instead of the expertise collection P. Given that the equivalence relation R_P corresponding to Π_P can be understood as an epistemic accessibility relation (by Theorem 4.3.3), we can interpret Π_P as expressing an *indistinguishability relation* over states: two states lie in the same partition cell if the source lacks expertise to distinguish them.

— 5 —

Belief Change with Non-Expert Sources

In the previous chapter we introduced a logical framework to reason about the expertise of sources and soundness of information. We now build on this framework to study a belief change problem in which expertise is not fixed upfront, but is to be estimated on the basis of reports from multiple sources. In this way we develop a logic-based analogue of the truth discovery aggregation problem, which complements the social-choice-style framework of Chapter 2. Specifically, we identify trustworthiness with *belief in expertise*: an agent deems source i trustworthy on φ if its belief set contains $\mathsf{E}_i\varphi$. By also including propositional formulas in belief sets, we are able to express beliefs about the actual world, i.e. the aggregation of the reports from sources.

Our point of departure is logic-based belief change in the AGM paradigm (named after Alchourrón, Gärdenfors, and Makinson (1985); refer back to Section 1.2 for an overview). To illustrate the problem – and how it differs from existing forms of belief change – consider the following scenario in a hospital. Suppose we observe the results of a blood test of patient A, confirming condition p. Assuming the test is reliable, AGM revision tells us how to revise our beliefs in light of the new information. Dr. X then claims that patient B suffers from the same condition, but Dr. Y disagrees. Given that doctors specialise in different areas and may make mistakes, who should we trust? Since the **Success** postulate ($\alpha \in K * \alpha$) assumes information is reliable, we are outside the realm of AGM revision, and must instead apply some form of *non-prioritised* revision (Hansson 1999a).

Suppose it now emerges that X had earlier claimed A did *not* suffer from condition p, contrary to the test results. We now have reason to suspect X may *lack expertise* on diagnosing p, and may subsequently revise beliefs about X's domain of expertise and the status of patient B (e.g. by opting to trust Y instead).

While simple, this example illustrates the key features of the belief change problem we study: we consider multiple sources, whose expertise is *a priori* unknown, providing reports on various instances of a problem domain. On the basis of these reports we form beliefs both about the expertise of the sources and the state of the world in each instance.

By including a distinguished *completely reliable* source (the test results in the example) we extend AGM revision. This is also analogous to *semi-supervised truth discovery* (Yin and Tan 2011; Rekatsinas et al. 2017), in which some ground truths are known ahead of time. In some respects we also extend approaches to non-prioritised revision (e.g. selective revision (Fermé and Hansson 1999), credibility-limited revision (Hansson et al. 2001), and trust-based revision (Booth and Hunter 2018)), which assume information about the reliability of sources is known up front. The problem is also related to *belief merging* (Konieczny and Pino Pérez 2002) which deals with combining belief bases from multiple sources; a detailed comparison will be given in Section 5.6.

Our work is also connected to trust and belief revision. As Yasser and Ismail (2020) note in recent work, trust and belief are inexorably linked. As set out in Chapter 2 in the context of truth discovery, we should accept reports from sources we believe are trustworthy, and we should trust sources whose reports turn out to be reliable. Trust and belief should also be revised in tandem, so that we may increase or decrease trust in a source as more reports are received, and revoke or reinstate previous reports from a source as its perceived trustworthiness changes.

To unify the trust and belief aspects, we work in (a fragment of) the language of expertise and soundness from Chapter 4, including formulas of the form $\mathsf{E}_i\varphi$, read as "source i has expertise on φ", and $\mathsf{S}_i\varphi$, read as "φ is sound for source i to report". The output of our belief change problem is then a collection of belief and knowledge sets in the language, describing what we *know* and *believe* about the expertise of the sources and the state of the world in each instance. For example, we should *know* reports from the reliable source are true, whereas reports from ordinary sources may only be believed.

We do, however, make some simplifying assumptions compared to the modal framework in Chapter 4. Firstly, we consider only a finite set of propositional variables, and identify states of the world with propositional valuations. Secondly, we assume expertise is closed under both intersections and complements, so that – by Theorem 4.3.2 – expertise of a source is fully captured by an equivalence relation; or equivalently, a *partition* of the propositional valuations. In other words, each source has a *indistinguishibility relation* over valuations, whereby any two valuations in the same partition cell are indistinguishable.

The semantics of expertise and soundness can be expressed directly in terms of partitions; we have that a source is an expert on a proposition φ exactly when they can distinguish every φ valuation from every $\neg\varphi$ valuation, and φ is sound for i if the "actual" state of the world is indistinguishable from a φ valuation.

We then make the assumption that *sources only report sound propositions*. That is, reports are only false due to sources overstepping the bounds of their expertise. In particular, we assume sources are honest in their reports, and that experts are always right.

Note that in our introductory example, the fact that we had a report from Dr. X on patient A (together with reliable information on patient A)

was essential for determining the expertise of X, and subsequently the status of patient B. While the patients are independent, reports on one can cause beliefs about the other to change, as we update our beliefs about the expertise of the sources.

In general we consider an arbitrary number of *cases*, which are seen as labels for instances of the domain. For example, a crowdsourcing worker may label multiple images, or a weather forecaster may give predictions for different locations. Each report in the input to the problem then refers to a specific case. Via these cases and the presence of the completely reliable source, we are able to model scenarios where some "ground truth" is available, listing how often sources have been correct/incorrect on a proposition (e.g. the *report histories* of Hunter (2021)). We can also generalise this scenario, e.g. by having only partial information about "previous" cases.

Throughout the chapter we make the assumption that *expertise is fixed across cases*: the expertise of a source does not depend on the particular instance of the domain we look at. For instance, the expertise of Dr. X is the same for patient A as for patient B. This is a simplifying assumption, and may rule out certain interpretations of the cases (e.g. if cases represent different points in time, it would be natural to let expertise evolve over time as per the dynamics of Section 4.6).

Contributions. The main contribution of this chapter is the formulation of a belief change problem in the setting of the logic of expertise developed in the previous chapter. This allows us to explore how belief and expertise-based trust should interact and evolve as reports are received from the various sources. We put forward several postulates and two concrete classes of operators – with a representation result for one class – and analyse these operators with respect to the postulates.

This chapter is an extension of Singleton and Booth (2022b). New material includes the results of Section 5.5.2 and the discussion in Section 5.6.

5.1 The Framework

Let $\mathcal{S}$ be a finite set of information sources. For convenience, we assume there is a *completely reliable* source in $\mathcal{S}$, which in a nod to the AGM tradition we denote by $*$. For example, we can treat our first-hand observations as if they are reported by $*$. Other sources besides $*$ will be termed *ordinary sources*. Let $\mathcal{C}$ be a finite set of *cases*, which we interpret as labels for different instances of the problem domain.

Syntax. In this chapter we work with the fragment of the language from Chapter 4, in which expertise and soundness formulas are restricted to propositional formulas only[1] and the universal modality A is excluded. Concretely, we assume a fixed *finite* set Prop of propositional variables, and let $\mathcal{L}_0$ denote the set of propositional formulas generated from Prop using the usual propositional connectives. Formulas in $\mathcal{L}_0$ are used to describe the "ontic" facts of the world in each case $c \in \mathcal{C}$. We use lower case Greek letters (φ, ψ etc) for formulas in $\mathcal{L}_0$. The classical logical consequence operator will be denoted by Cn_0, and $\equiv$ denotes equivalence of propositional formulas.

The extended language of expertise $\mathcal{L}$ additionally describes the expertise of the sources, and is defined by the following grammar:

$$\Phi ::= p \mid \Phi \wedge \Phi \mid \neg \Phi \mid \mathsf{E}_i \varphi \mid \mathsf{S}_i \varphi$$

where $i \in \mathcal{S}$, $p \in \mathsf{Prop}$ and $\varphi \in \mathcal{L}_0$. We introduce Boolean connectives $\vee$, $\rightarrow$, $\leftrightarrow$ and $\bot$ as abbreviations. We use upper case Greek letters (Φ, Ψ etc) for formulas in $\mathcal{L}$. For $\Gamma \subseteq \mathcal{L}$, we write $[\Gamma] = \Gamma \cap \mathcal{L}_0$ for the propositional formulas in Γ.

As before, the intuitive reading of $\mathsf{E}_i\varphi$ is "source i has expertise on φ". The intuitive reading of $\mathsf{S}_i\varphi$ is that "φ sound for i to report": i.e. that φ is true up to the expertise of i. In other words, the parts of φ on which i has expertise are true. Since both operators are restricted to propositional formulas, we will not consider iterated formulas such as $\mathsf{E}_i\mathsf{S}_j\varphi$.

Semantics. The semantics in this chapter are, in essence, a special case of Chapter 4. Instead of considering arbitrary sets of possible states, as in Definition 4.1.1, we fix states as propositional valuations over Prop. Expertise is also assumed to be closed under both intersections and complements for all sources. At the same time, we generalise slightly by considering the distinguished source $*$ and multiple cases $c \in \mathcal{C}$. For convenience we also offer a different presentation, using *partitions* instead of expertise collections to represent expertise.

Formally, let $\mathcal{V}$ denote the set of propositional valuations over Prop. For each $\varphi \in \mathcal{L}_0$, the set of valuations making φ true is denoted by $\|\varphi\|$. A *world* $W = \langle \{v_c\}_{c \in \mathcal{C}}, \{\Pi_i\}_{i \in \mathcal{S}} \rangle$ is a possible complete specification of the environment we find ourselves in:

- $v_c \in \mathcal{V}$ is the "true" valuation at case $c \in \mathcal{C}$;
- Π_i is a partition of $\mathcal{V}$ for each $i \in \mathcal{S}$, representing the "true" expertise of source i; and
- Π_* is the unit partition $\{\{v\} \mid v \in \mathcal{V}\}$.

[1]While this assumption is made for simplicity's sake, we do not lose much by excluding iterated applications of E and S, at least for expertise models closed under intersections and complements. Indeed, we have that $\mathsf{E}\varphi$ either holds globally in a model or holds nowhere, so $\mathsf{EE}\varphi$ always holds. One can show that $\mathsf{ES}\psi$ also always holds in such models, by taking $\varphi = \mathsf{S}\varphi$ in Proposition 4.2.1 (1) and recalling that $\mathsf{SS}\psi \rightarrow \mathsf{S}\psi$ is valid. Similarly, one can show that $\mathsf{SS}\varphi \leftrightarrow \mathsf{S}\varphi$ and $\mathsf{SE}\varphi \leftrightarrow \mathsf{E}\varphi$ in such models.

Let $\mathcal{W}$ denote the set of all worlds. Note that the partition corresponding to the distinguished source $*$ is fixed in all worlds as the finest possible partition, reflecting the fact that $*$ is completely reliable.

For any partition Π and valuation v, write $\Pi[v]$ for the unique cell in Π containing v. For a set of valuations U, write $\Pi[U] = \bigcup_{v \in U} \Pi[v]$. For brevity, we write $\Pi[\varphi]$ for $\Pi[\|\varphi\|]$. Then $\Pi[\varphi]$ is the set of valuations indistinguishable from a φ valuation.

For our belief change problem we will be interested in maintaining a collection of several belief sets, describing beliefs about each case $c \in \mathcal{C}$. Towards determining when a world W models such a collection, we define semantics for $\mathcal{L}$ formulas with respect to a world and a case:

$$\begin{aligned} W, c \models p &\iff v_c \in \|p\| \\ W, c \models \mathsf{E}_i\varphi &\iff \Pi_i[\varphi] = \|\varphi\| \\ W, c \models \mathsf{S}_i\varphi &\iff v_c \in \Pi_i[\varphi] \end{aligned}$$

where $i \in \mathcal{S}$, $\varphi \in \mathcal{L}_0$, and the clauses for conjunction and negation are the expected ones. Since $\|\varphi\| \subseteq \Pi_i[\varphi]$ always holds, we have that $\mathsf{E}_i\varphi$ holds iff there is no $\neg\varphi$ valuation which is indistinguishable from a φ valuation (c.f. Booth and Hunter (2018)). Note that since each source i has only a single partition Π_i used to interpret the expertise formulas, the truth value of $\mathsf{E}_i\varphi$ does not depend on the case c. On the other hand, $\mathsf{S}_i\varphi$ holds in case c iff the c-valuation of W is indistinguishable from some model of φ. That is, it is consistent with i's expertise that φ is true.

Note that the mapping $2^{\mathcal{V}} \to 2^{\mathcal{V}}$ given by $U \mapsto \Pi[U]$ satisfies the *Kuratowski closure axioms*,[2] so can be considered a closure operator of the set of propositional valuations. Then $W, c \models \mathsf{E}_i\varphi$ iff $\|\varphi\|$ is closed in $\mathcal{V}$, and $W, c \models \mathsf{S}_i\varphi$ iff v_c lies in the closure of $\|\varphi\|$, i.e. φ is true after closing $\|\varphi\|$ along the lines of the expertise of source i. Also note that $\Pi[U] = U$ iff U can be expressed as a union of the partition cells in Π, so that $W, c \models \mathsf{E}_i\varphi$ can alternatively be interpreted as saying φ is a disjunction of stronger formulas on which i also has expertise.

Also note that if φ is a propositional tautology, $\mathsf{E}_i\varphi$ holds for every source i. Thus, all sources are experts on *something*, even if just the tautologies.

The semantics so defined are indeed the same as those of Chapter 4, as the following result shows.

Proposition 5.1.1. *Let W be a world. Then there is a multi-source expertise model $M = (X, \{P_i\}_{i \in \mathcal{S}}, V)$ and $\{x_c\}_{c \in \mathcal{C}} \subseteq X$ such that for all $\Phi \in \mathcal{L}$ and $c \in \mathcal{C}$,*

$$W, c \models \Phi \iff M, x_c \models \Phi. \tag{5.1}$$

Moreover, (i) X is finite; (ii) each P_i is closed under intersections and complements; and (iii) using the notation from Definition 4.3.2, $uR_{P_i}v$ iff $\Pi_i[u] = \Pi_i[v]$, i.e. R_{P_i} is the equivalence relation associated with the partition Π_i.

[2] That is, (i) $\Pi[\emptyset] = \emptyset$, (ii) $U \subseteq \Pi[U]$, (iii) $\Pi[\Pi[U]] = \Pi[U]$, and (iv) $\Pi[U_1 \cup U_2] = \Pi[U_1] \cup \Pi[U_2]$.

Proof. Take $X = \mathcal{V}$ and set $V(p) = \|p\|$. For each $i \in \mathcal{S}$, set $P_i = \{A \subseteq \mathcal{V} \mid \Pi_i[A] = A\}$. For each $c \in \mathcal{C}$, simply let $x_c = v_c$. Then one can easily show that for all $U \subseteq \mathcal{V}$ and $i \in \mathcal{S}$,

$$\Pi_i[U] = \bigcap\{A \in P_i \mid U \subseteq A\}. \tag{5.2}$$

A simple induction on $\mathcal{L}$ formulas then shows (5.1).

Since Prop is finite there are only finitely many propositional valuations, and thus $X = \mathcal{V}$ is finite. It is easily checked that each P_i is closed under intersections and complements using properties of partitions. Finally, we have by (5.2) and the definition of R_{P_i} that

$$\begin{aligned} uR_{P_i}v &\iff \forall A \in P_i : (v \in A \implies u \in A) \\ &\iff u \in \bigcap\{A \in P_i \mid v \in A\} \\ &\iff u \in \bigcap\{A \in P_i \mid \{v\} \subseteq A\} \\ &\iff u \in \Pi_i[\{v\}] \\ &\iff \Pi_i[u] = \Pi_i[v] \end{aligned}$$

as required. □

In other words, a world corresponds to a particular kind of expertise model together with a state x_c for each case $c \in \mathcal{C}$. Having shown this equivalence, we henceforth deal exclusively with worlds and models instead of expertise models and collections. We come to an example.

Example 5.1.1. *Let us extend the hospital example from the introduction. Let $\mathcal{S} = \{*, \mathsf{X}, \mathsf{Y}\}$ denote the reliable source, Dr. X and Dr. Y, and let $\mathcal{C} = \{A, B\}$ denote patients A and B. Consider propositional variables $\mathsf{Prop} = \{p, q\}$, standing for conditions p and q respectively. Suppose that X has expertise on diagnosing condition q only, whereas Y only has expertise on p. For the sake of the example, suppose that patient A suffers from both conditions, and patient B suffers only from condition q. This situation is modelled by the following world $W = \langle \{v_c\}_{c\in\{A,B\}}, \{\Pi_i\}_{i\in\{*,\mathsf{X},\mathsf{Y}\}}\rangle$:*

$$\begin{array}{rr} v_A = pq; & v_B = \bar{p}q; \\ \Pi_{\mathsf{X}} = pq, \bar{p}q \mid p\bar{q}, \bar{p}\bar{q}; & \Pi_{\mathsf{Y}} = pq, p\bar{q} \mid \bar{p}q, \bar{p}\bar{q}, \end{array}$$

where Π_* is the unit partition. This world is also depicted graphically in Fig. 5.1. We have $W, c \models \mathsf{E}_{\mathsf{X}}q \wedge \mathsf{E}_{\mathsf{Y}}p$ for each $c \in \{A, B\}$. Also note that $W, A \models p$ (A suffers from p), $W, A \models \mathsf{S}_{\mathsf{X}}\neg p$ (it is sound for X to report otherwise; this holds since $\Pi_{\mathsf{X}}[\neg p] = \{pq, \bar{p}q\} \cup \{p\bar{q}, \bar{p}\bar{q}\} \ni pq = v_A$), but $W, A \models \neg\mathsf{S}_{\mathsf{Y}}\neg p$ (the same formula is not *sound for Y; we have $\Pi_{\mathsf{Y}}[\neg p] = \{\bar{p}q, \bar{p}\bar{q}\} = \|\neg p\| \not\ni pq = v_A$).*

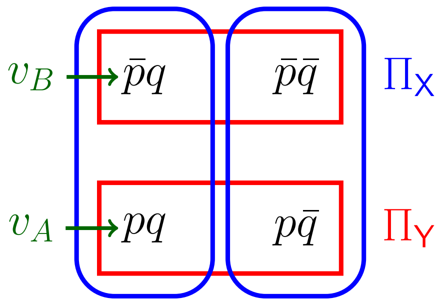


Figure 5.1: Depiction of the world W defined in Example 5.1.1.

Say Φ is *valid* if $W, c \models \Phi$ for all $W \in \mathcal{W}$ and $c \in \mathcal{C}$. For future reference we collect a list of validities.[3]

Proposition 5.1.2. *For any $i \in \mathcal{S}$, $c \in \mathcal{C}$ and $\varphi, \psi \in \mathcal{L}_0$, the following formulas are valid*

1. $\mathsf{S}_i\varphi \leftrightarrow \mathsf{S}_i\psi$ *and* $\mathsf{E}_i\varphi \leftrightarrow \mathsf{E}_i\psi$*, whenever* $\varphi \equiv \psi$
2. $\mathsf{E}_i\varphi \leftrightarrow \mathsf{E}_i\neg\varphi$ *and* $\mathsf{E}_i\varphi \wedge \mathsf{E}_i\psi \rightarrow \mathsf{E}_i(\varphi \wedge \psi)$
3. $\mathsf{E}_ip_1 \wedge \cdots \wedge \mathsf{E}_ip_k \rightarrow \mathsf{E}_i\varphi$*, where* $p_1, \ldots, p_k$ *are the propositional variables appearing in* φ
4. $\mathsf{E}_i\varphi \wedge \mathsf{S}_i\varphi \rightarrow \varphi$*, and* $\mathsf{S}_i\varphi \wedge \neg\varphi \rightarrow \neg\mathsf{E}_i\varphi$
5. $\mathsf{S}_i\varphi \wedge \mathsf{S}_i\neg\varphi \rightarrow \neg\mathsf{E}_i\varphi$
6. $\mathsf{S}_*\varphi \leftrightarrow \varphi$ *and* $\mathsf{E}_*\varphi$

We comment on each property before giving the proof. (1) states syntax-irrelevance properties. (2) says that expertise is symmetric with respect to negation, and closed under conjunctions. Intuitively, symmetry means that i is an expert on φ if they know *whether or not* φ holds. (3) says that expertise on each propositional variable in φ is sufficient for expertise on φ itself. (4) says that, in the presence of expertise, soundness of φ is sufficient for φ to in fact be true. (5) says that if both φ and $\neg\varphi$ are true up to the expertise of i, then i cannot have expertise on φ. Finally, (6) says that the reliable source $*$ has expertise on *all* formulas, and thus φ is sound for $*$ iff it is true.

Proof.

1. If $\varphi \equiv \psi$ then $\|\varphi\| = \|\psi\|$; since the semantics for $\mathsf{S}_i\varphi$ and $\mathsf{E}_i\varphi$ only refer to $\|\varphi\|$ (and likewise for ψ), we have that $\mathsf{S}_i\varphi \leftrightarrow \mathsf{S}_i\psi$ and $\mathsf{E}_i\varphi \leftrightarrow \mathsf{E}_i\psi$ are valid.

[3] Note that some of these validities follow from Proposition 5.1.1 and the validities in Chapter 4.

2. For the first validity, suppose $W, c \models \mathsf{E}_i\varphi$. Then $\|\varphi\| = \Pi_i[\varphi]$. We show $W, c \models \mathsf{E}_i\neg\varphi$. Indeed, take $v \in \Pi_i[\neg\varphi]$. Then there is $v' \in \|\neg\varphi\|$ such that $v \in \Pi_i[v']$. Thus $v' \in \Pi_i[v]$ also. Supposing for contradiction that $v \in \|\varphi\|$, we get
$$v' \in \Pi_i[v] \subseteq \Pi_i[\varphi] = \|\varphi\|.$$
But then $v' \in \|\neg\varphi\| \cap \|\varphi\| = \emptyset$; contradiction. Hence $v \notin \|\varphi\|$, i.e. $v \in \|\neg\varphi\|$. This shows that $\Pi_i[\neg\varphi] \subseteq \|\neg\varphi\|$, so $W, c \models \mathsf{E}_i\neg\varphi$.

   We have shown that $\mathsf{E}_i\varphi \to \mathsf{E}_i\neg\varphi$ is valid. For the converse note that, by symmetry, $\mathsf{E}_i\neg\varphi \to \mathsf{E}_i\neg\neg\varphi$ is valid; since $\mathsf{E}_i\neg\neg\varphi$ is equivalent to $\mathsf{E}_i\varphi$ by (1) we get $\mathsf{E}_i\varphi \leftrightarrow \mathsf{E}_i\neg\varphi$.

   For the second validity, suppose $W, c \models \mathsf{E}_i\varphi \wedge \mathsf{E}_i\psi$. Note that
$$\Pi_i[\varphi \wedge \psi] \subseteq \Pi_i[\varphi] = \|\varphi\|$$
and, similarly, $\Pi_i[\varphi \wedge \psi] \subseteq \|\psi\|$. Hence
$$\Pi_i[\varphi \wedge \psi] \subseteq \|\varphi\| \cap \|\psi\| = \|\varphi \wedge \psi\|,$$
which shows $W, c \models \mathsf{E}_i(\varphi \wedge \psi)$.

3. Let φ be a propositional formula, and let $p_1, \ldots, p_k$ be the variables appearing in φ. Let $\widehat{\mathcal{L}_0} \subseteq \mathcal{L}_0$ be the propositional formulas over $p_1, \ldots p_k$ generated only using conjunction and negation. Then there is some $\psi \in \widehat{\mathcal{L}_0}$ with $\varphi \equiv \psi$.

   Suppose $W, c \models \mathsf{E}_i p_1 \wedge \cdots \wedge \mathsf{E}_i p_k$. By this assumption and the properties in (2), one can show by induction that $W, c \models \mathsf{E}_i\theta$ for all $\theta \in \widehat{\mathcal{L}_0}$. In particular, $W, c \models \mathsf{E}_i\psi$. Since $\varphi \equiv \psi$, we get $W, c \models \mathsf{E}_i\varphi$.

4. Suppose $W, c \models \mathsf{E}_i\varphi \wedge \mathsf{S}_i\varphi$. Then $v_c \in \Pi_i[\varphi] = \|\varphi\|$, so $W, c \models \varphi$. Hence $\mathsf{E}_i\varphi \wedge \mathsf{S}_i\varphi \to \varphi$ is valid. Similarly, $\mathsf{S}_i\varphi \wedge \neg\varphi \to \neg\mathsf{E}_i\varphi$ is valid.

5. Suppose $W, c \models \mathsf{S}_i\varphi \wedge \mathsf{S}_i\neg\varphi$, and, for contradiction, $W, c \models \mathsf{E}_i\varphi$. On the one hand we have $W, c \models \mathsf{E}_i\varphi \wedge \mathsf{S}_i\varphi$, so (4) gives $W, c \models \varphi$. On the other hand, $W, c \models \mathsf{E}_i\varphi$ gives $W, c \models \mathsf{E}_i\neg\varphi$ by (2), so $W, c \models \mathsf{E}_i\neg\varphi \wedge \mathsf{S}_i\neg\varphi$; by (4) again we have $W, c \models \neg\varphi$. But then $W, c \models \varphi \wedge \neg\varphi$ – contradiction.

6. Since the distinguished source $*$ has the unit partition Π_* in any world W, we have $\Pi_*[\varphi] = \|\varphi\|$, so $W, c \models \mathsf{E}_*\varphi$. Similarly, $W, c \models \mathsf{S}_i\varphi$ iff $v_c \in \Pi_*[\varphi] = \|\varphi\|$ iff $W, c \models \varphi$.

□

Case-Indexed Collections. In the remainder of this chapter we will be interested in forming beliefs about each case $c \in \mathcal{C}$. To do so we use collections of belief sets $G = \{\Gamma_c\}_{c\in\mathcal{C}}$, with $\Gamma_c \subseteq \mathcal{L}$, indexed by cases. Say a world W is a *model* of G iff

$$W, c \models \Phi \text{ for all } c \in \mathcal{C} \text{ and } \Phi \in \Gamma_c,$$

i.e. iff W satisfies all formulas in G in the relevant case. Let $\mathrm{mod}(G)$ denote the models of G, and say that G is *consistent* if $\mathrm{mod}(G) \neq \emptyset$. For $c \in \mathcal{C}$, define the *c-consequences*

$$\mathrm{Cn}_c(G) = \{\Phi \in \mathcal{L} \mid \forall W \in \mathrm{mod}(G), W, c \models \Phi\}.$$

That is, Φ is a c-consequence of a collection G if Φ holds in case c for every model W of G. We write $\mathrm{Cn}(G)$ for the collection $\{\mathrm{Cn}_c(G)\}_{c\in\mathcal{C}}$.

Example 5.1.2. *Suppose $\mathcal{C} = \{c_1, c_2, c_3\}$, and define G by $\Gamma_{c_1} = \{\mathsf{S}_i(p \wedge q)\}$, $\Gamma_{c_2} = \{\mathsf{E}_i p\}$ and $\Gamma_{c_3} = \{\mathsf{E}_i q\}$. Then, since expertise holds independently of case, any model W of G has $W, c_1 \models \mathsf{E}_i p \wedge \mathsf{E}_i q$. By Proposition 5.1.2 part (3), $W, c_1 \models \mathsf{E}_i(p \wedge q)$. Since W satisfies Γ_{c_1} in case c_1, Proposition 5.1.2 part (4) gives $W, c_1 \models p \wedge q$. Since W was an arbitrary model of G, we have $p \wedge q \in \mathrm{Cn}_{c_1}(G)$, i.e. $p \wedge q$ is a c_1-consequence of G. This illustrates how information about distinct cases can be brought together to have consequences for other cases.*

For two collections $G = \{\Gamma_c\}_{c\in\mathcal{C}}$, $D = \{\Delta_c\}_{c\in\mathcal{C}}$, write $G \sqsubseteq D$ iff $\Gamma_c \subseteq \Delta_c$ for all c, and let $G \sqcup D$ denote the collection $\{\Gamma_c \cup \Delta_c\}_{c\in\mathcal{C}}$. With this notation, the case-indexed consequence operator satisfies analogues of the Tarskian consequence properties.[4]

Say a collection G is *closed* if $\mathrm{Cn}(G) = G$. Closed collections provide an idealised representation of beliefs, which will become useful later on. For instance, when G is closed any expertise formula $\mathsf{E}_i\varphi$ is either present in Γ_c for all cases c or for none, reflecting the fact that expertise is fixed across cases. We also have that propositional beliefs about the world in each case c are closed under classical consequences.

Proposition 5.1.3. *Suppose a collection $G = \{\Gamma_c\}_{c\in\mathcal{C}}$ is closed. Then*

1. $\mathsf{E}_i\varphi \in \Gamma_c$ *if and only if* $\mathsf{E}_i\varphi \in \Gamma_d$.
2. $\mathrm{Cn}_0\,[\Gamma_c] = [\Gamma_c]$.

Proof.

1. Suppose $\mathsf{E}_i\varphi \in \Gamma_c$. Since $G = \mathrm{Cn}(G)$ we have $\Gamma_d = \mathrm{Cn}_d(G)$. Take $W \in \mathrm{mod}(G)$. Then W satisfies all formulas of Γ_c in case c, so $W, c \models \mathsf{E}_i\varphi$. As expertise is independent of case, this means $W, d \models \mathsf{E}_i\varphi$. By definition

[4] That is, (i) $G \sqsubseteq \mathrm{Cn}(G)$, (ii) $G \sqsubseteq D$ implies $\mathrm{Cn}(G) \sqsubseteq \mathrm{Cn}(D)$, and (iii) $\mathrm{Cn}(\mathrm{Cn}(G)) = \mathrm{Cn}(G)$.

of d-consequences, we have $\mathsf{E}_i\varphi \in \mathrm{Cn}_d(G) = \Gamma_d$ as required. The reverse direction holds by symmetry.

2. Clearly $[\Gamma_c] \subseteq \mathrm{Cn}_0\,[\Gamma_c]$. For the reverse inclusion, suppose $\varphi \in \mathrm{Cn}_0\,[\Gamma_c]$. We will show $\varphi \in \mathrm{Cn}_c(G)$, which implies $\varphi \in [\Gamma_c]$ since G is closed. Indeed, take $W \in \mathrm{mod}(G)$. Then, $W, c \models \psi$ for all $\psi \in [\Gamma]_c$. By definition of the semantics for propositional formulas, this means the valuation v_c^W satisfies each ψ, so v_c^W is a model of $[\Gamma_c]$. By the definition of the classical consequence operator Cn_0, this means v_c^W is also a model of φ, i.e. $W, c \models \varphi$. Hence $\varphi \in \mathrm{Cn}_c(G)$ as required.

□

In propositional logic, $\|\cdot\|$ is a 1-to-1 correspondence between closed sets of formulas and sets of valuations. This is not so in our setting, since some subsets of $\mathcal{W}$ do not arise as the models of any collection. Instead, we have a 1-to-1 correspondence into a restricted collection of sets of worlds. Borrowing the terminology of Delgrande, Peppas, and Woltran (2018), say a set of worlds $S \subseteq \mathcal{W}$ is *elementary* if $S = \mathrm{mod}(G)$ for some collection $G = \{\Gamma_c\}_{c\in\mathcal{C}}$.[5]

Elementariness is characterised by a certain closure condition. Say that two worlds W, W' are *partition-equivalent* if $\Pi_i^W = \Pi_i^{W'}$ for all sources i, and say W is a *valuation combination* from a set $S \subseteq \mathcal{W}$ if for all cases c there is $W_c \in S$ such that $v_c^W = v_c^{W_c}$. Then a set is elementary iff it is closed under valuation combinations of partition-equivalent worlds.

Proposition 5.1.4. *$S \subseteq \mathcal{W}$ is elementary if and only if the following condition holds: for all $W \in \mathcal{W}$ and $W_1, W_2 \in S$, if W is partition-equivalent to both W_1, W_2 and W is a valuation combination from $\{W_1, W_2\}$, then $W \in S$.*

Proof. “if”: Suppose the stated condition holds for $S \subseteq \mathcal{W}$. Form a collection $G = \{\Gamma_c\}_{c\in\mathcal{C}}$ by setting $\Gamma_c = \{\Phi \in \mathcal{L} \mid S \subseteq \mathrm{mod}_c(\Phi)\}$. Clearly $S \subseteq \mathrm{mod}(G)$. For the reverse inclusion, suppose $W \in \mathrm{mod}(G)$. For any set of valuations $U \subseteq \mathcal{V}$, let φ_U be any propositional formula with $\|\varphi_U\| = U$. For each $c \in \mathcal{C}$, consider the formula

$$\Phi_c = \bigvee_{W'\in S} \left(\varphi_{\{v_c^{W'}\}} \wedge \bigwedge_{i\in\mathcal{S}} \bigwedge_{U\subseteq\mathcal{V}} R_{W',i,U} \right)$$

where

$$R_{W',i,U} = \begin{cases} \mathsf{E}_i\varphi_U, & W', c_0 \models \mathsf{E}_i\varphi_U \\ \neg\mathsf{E}_i\varphi_U, & \text{otherwise} \end{cases}$$

for some fixed case $c_0 \in \mathcal{C}$. It is straightforward to see that each $W' \in S$ satisfies its corresponding disjunct at case c, so $\Phi_c \in \Gamma_c$. Hence $W \in \mathrm{mod}(G)$

[5]Non-elementary sets can also exist for weaker logics (such as Horn logic (Delgrande, Peppas, and Woltran 2018)) which lack the syntactic expressivity to identify all sets of models. In our framework, $\mathcal{C}$-indexed collections are not expressive enough to specify *combinations of valuations*, since each Γ_c only says something about the valuation for c.

implies $W, c \models \Phi_c$ for each c. Consequently, for each c there is some $W_c \in S$ such that (i) $v_c^W = v_c^{W_c}$; and (ii) for each $i \in \mathcal{S}$ and $U \subseteq \mathcal{V}$, $W, c \models \mathsf{E}_i\varphi_U$ iff $W_c, c \models \mathsf{E}_i\varphi_U$. From (i), W is a valuation combination from $\{W_c\}_{c\in\mathcal{C}}$. From (ii) it can be shown that in fact $\Pi_i^W = \Pi_i^{W_c}$ for each c and i; that is, W is partition-equivalent to each W_c. In particular, all the W_c are partition-equivalent to each other. Repeatedly applying the closure condition assumed to hold for S, we see that $W \in S$ as required.

"only if": Suppose S is elementary, i.e. $S = \mathrm{mod}(G)$ for some collection $G = \{\Gamma_c\}_{c\in\mathcal{C}}$, and let W, W_1, W_2 be as in the statement of the proposition. Take $c \in \mathcal{C}$ and $\Phi \in \Gamma_c$. We will show $W, c \models \Phi$. By assumption, there is $n \in \{1, 2\}$ such that $v_c^W = v_c^{W_n}$. It can be shown by induction on $\mathcal{L}$ formulas that, since W and W_n are partition-equivalent and have the same c valuation, $W, c \models \Phi$ iff $W_n, c \models \Phi$. But $W_n \in S = \mathrm{mod}(G)$ implies $W_n, c \models \Phi$, so $W, c \models \Phi$ too. Since $\Phi \in \Gamma_c$ was arbitrary, we have $W \in \mathrm{mod}(G) = S$ as required. □

5.2 The Problem

With the framework set out, we can formally define the problem. As input, we consider sequences of *reports*, typically denoted σ, where each report is a triple $\langle i, c, \varphi\rangle \in \mathcal{S} \times \mathcal{C} \times \mathcal{L}_0$ with $\varphi \not\equiv \bot$. Such a report represents that *source i reports φ to hold in case c*. Note that we only allow sources to make *propositional* reports. A *belief and expertise operator* produces a collection of belief and knowledge sets on the basis of such sequences.

Definition 5.2.1. *A* belief and expertise operator *maps each sequence of reports σ to a pair $\langle B^\sigma, K^\sigma\rangle$, where $B^\sigma = \{B_c^\sigma\}_{c\in\mathcal{C}}$ is a collection of* belief sets *$B_c^\sigma \subseteq \mathcal{L}$ and $K^\sigma = \{K_c^\sigma\}_{c\in\mathcal{C}}$ is a collection of* knowledge sets *$K_c^\sigma \subseteq \mathcal{L}$.*

5.2.1 Basic Postulates

We immediately narrow the scope of operators under consideration by introducing some basic postulates which are expected to hold. In what follows, say a sequence σ is *$*$-consistent* if for each $c \in \mathcal{C}$ the set $\{\varphi \mid \langle *, c, \varphi\rangle \in \sigma\} \subseteq \mathcal{L}_0$ is classically consistent. Write G_{snd}^σ for the collection with $(G_{\mathsf{snd}}^\sigma)_c = \{\mathsf{S}_i\varphi \mid \langle i, c, \varphi\rangle \in \sigma\}$, i.e. the collection of soundness statements corresponding to the reports in σ.

Closure. $B^\sigma = \mathrm{Cn}(B^\sigma)$ and $K^\sigma = \mathrm{Cn}(K^\sigma)$

Containment. $K^\sigma \sqsubseteq B^\sigma$

Consistency. If σ is $*$-consistent, B^σ and K^σ are consistent

Soundness. If $\langle i, c, \varphi\rangle \in \sigma$, then $\mathsf{S}_i\varphi \in K_c^\sigma$

K-bound. $K^\sigma \sqsubseteq \mathrm{Cn}(G_{\mathsf{snd}}^\sigma \sqcup K^\emptyset)$

Prior-extension. $K^{\emptyset} \sqsubseteq K^{\sigma}$

Rearrangement. If σ is a permutation of ρ, then $B^{\sigma} = B^{\rho}$ and $K^{\sigma} = K^{\rho}$

Equivalence. If $\varphi \equiv \psi$ then $B^{\sigma\cdot\langle i,c,\varphi\rangle} = B^{\sigma\cdot\langle i,c,\psi\rangle}$ and $K^{\sigma\cdot\langle i,c,\varphi\rangle} = K^{\sigma\cdot\langle i,c,\psi\rangle}$

Closure says that the belief and knowledge collections are closed under logical consequence. In light of earlier remarks, this implies that the propositional belief sets $[B^{\sigma}_c]$ are closed under (propositional) consequence, and that $\mathsf{E}_i\varphi \in B^{\sigma}_c$ iff $\mathsf{E}_i\varphi \in B^{\sigma}_d$. **Containment** says that everything which is known is also believed. **Consistency** ensures the output is always consistent, provided we are not in the degenerate case where $*$ gives inconsistent reports. **Soundness** says we *know* that all reports are sound in their respective cases. This formalises our assumption that sources are *honest*, i.e. that false reports only arise due to lack of expertise. By Proposition 5.1.2 part (4) it also implies *experts are always right*: if a source has expertise on their report then it must be true. While **Soundness** places a lower bound on knowledge, **K-bound** places an upper bound: knowledge cannot go beyond the soundness statements corresponding to the reports in σ together with the prior knowledge $K^{\emptyset}$. That is, from the point view of knowledge, a new report of $\langle i, c, \varphi\rangle$ only allows us to learn $\mathsf{S}_i\varphi$ in case c (and to combine this with other reports and prior knowledge). Note that the analogous property for belief is *not* desirable: we want to be more liberal when it comes to beliefs, and allow for *defeasible inferences* going beyond the mere fact that reports are sound. **Prior-extension** says that knowledge after a sequence σ extends the prior knowledge on the empty sequence $\emptyset$. **Rearrangement** says that the order in which reports are received is irrelevant. This can be justified on the basis that we are reasoning about *static worlds* for each case c, so that there is no reason to see more "recent" reports as any more or less important or truthful than earlier ones.[6] Consequently, we can essentially view the input as a *multi-set* of belief sets – one for each source – bringing us close to the setting of belief merging. This postulate also appears as the commutativity postulate **(Com)** in the work of Schwind and Konieczny (2020). Finally, **Equivalence** says that the syntactic form of reports is irrelevant.

Taking all the basic postulates together, the knowledge component K^{σ} is fully determined once $K^{\emptyset}$ is chosen.

Proposition 5.2.1. *Suppose an operator satisfies the basic postulates. Then*

1. $K^{\sigma} = \mathrm{Cn}(G^{\sigma}_{\mathsf{snd}} \sqcup K^{\emptyset})$.
2. $K^{\emptyset} = \mathrm{Cn}(\emptyset)$ *iff* $K^{\sigma} = \mathrm{Cn}(G^{\sigma}_{\mathsf{snd}})$ *for all* σ.

Proof.

[6] This argument is from (Delgrande, Dubois, and Lang 2006).

1. The "$\sqsubseteq$" inclusion is just **K-bound**. For the "$\sqsupseteq$" inclusion, note that $G^{\sigma}_{\mathsf{snd}} \sqsubseteq K^{\sigma}$ by **Soundness**, and $K^{\emptyset} \sqsubseteq K^{\sigma}$ by **Prior-extension**. Hence

$$G^{\sigma}_{\mathsf{snd}} \sqcup K^{\emptyset} \sqsubseteq K^{\sigma}.$$

By monotonicity of Cn,

$$\mathrm{Cn}(G^{\sigma}_{\mathsf{snd}} \sqcup K^{\emptyset}) \sqsubseteq \mathrm{Cn}(K^{\sigma}) = K^{\sigma}$$

where we use **Closure** in the final step.

2. "if": Suppose $K^{\sigma} = \mathrm{Cn}(G^{\sigma}_{\mathsf{snd}})$ for all σ. Taking $\sigma = \emptyset$ we obtain

$$K^{\sigma} = \mathrm{Cn}(G^{\emptyset}_{\mathsf{snd}}) = \mathrm{Cn}(\emptyset).$$

"only if": Suppose $K^{\emptyset} = \mathrm{Cn}(\emptyset)$. Take any sequence σ. By **K-bound**,

$$K^{\sigma} \sqsubseteq \mathrm{Cn}(G^{\sigma}_{\mathsf{snd}} \sqcup \mathrm{Cn}(\emptyset)) = \mathrm{Cn}(G^{\sigma}_{\mathsf{snd}})$$

On the other hand, **Soundness** and **Closure** give $\mathrm{Cn}(G^{\sigma}_{\mathsf{snd}}) \sqsubseteq K^{\sigma}$. Hence $K^{\sigma} = \mathrm{Cn}(G^{\sigma}_{\mathsf{snd}})$.

□

The choice of $K^{\emptyset}$ depends on the scenario one wishes to model. While $\mathrm{Cn}(\emptyset)$ is a sensible choice if the sequence σ is all we have to go on, we allow $K^{\emptyset} \neq \mathrm{Cn}(\emptyset)$ in case *prior knowledge* is available (for example, the expertise of particular sources may be known ahead of time).

Another important property of knowledge, which follows from the basic postulates, says that *knowledge is monotonic*: knowledge after receiving σ and ρ together is just the case-wise union of K^{σ} and K^{ρ}.

K-conjunction. $K^{\sigma\cdot\rho} = \mathrm{Cn}(K^{\sigma} \sqcup K^{\rho})$

This postulate reflects the idea that one should be cautious when it comes to knowledge, a formula should only be accepted as known if it won't be given up in light of new information.

Proposition 5.2.2. *Any operator satisfying the basic postulates satisfies **K-conjunction**.*

Proof. Suppose an operator satisfies the basic postulates, and take sequences σ and ρ. By Proposition 5.2.1,

$$K^{\sigma\cdot\rho} = \mathrm{Cn}(G^{\sigma\cdot\rho}_{\mathsf{snd}} \sqcup K^{\emptyset})$$

Note that $G^{\sigma\cdot\rho}_{\mathsf{snd}} = G^{\sigma}_{\mathsf{snd}} \sqcup G^{\rho}_{\mathsf{snd}}$. Hence we may write

$$\begin{aligned} K^{\sigma\cdot\rho} &= \mathrm{Cn}(G^{\sigma}_{\mathsf{snd}} \sqcup G^{\rho}_{\mathsf{snd}} \sqcup K^{\emptyset}) \\ &= \mathrm{Cn}((G^{\sigma}_{\mathsf{snd}} \sqcup K^{\emptyset}) \sqcup (G^{\rho}_{\mathsf{snd}} \sqcup K^{\emptyset})) \end{aligned}$$

By Proposition 5.2.1 again, we have $K^\sigma = \mathrm{Cn}(G^\sigma_{\mathsf{snd}} \sqcup K^\emptyset)$ and $K^\rho = \mathrm{Cn}(G^\rho_{\mathsf{snd}} \sqcup K^\emptyset)$. It is easily verified that for any collections G, D, we have

$$\mathrm{Cn}(G \sqcup D) = \mathrm{Cn}(\mathrm{Cn}(G) \sqcup \mathrm{Cn}(D)).$$

Consequently,

$$\begin{aligned} K^{\sigma \cdot \rho} &= \mathrm{Cn}(\mathrm{Cn}(G^\sigma_{\mathsf{snd}} \sqcup K^\emptyset) \sqcup \mathrm{Cn}(G^\rho_{\mathsf{snd}} \sqcup K^\emptyset)) \\ &= \mathrm{Cn}(K^\sigma \sqcup K^\rho) \end{aligned}$$

as required for **K-conjunction**. □

The postulates also imply some useful properties linking *trust* (seen as belief in expertise) and *belief/knowledge*.

Proposition 5.2.3. *Suppose an operator satisfies the basic postulates. Then*

1. *If* $\varphi \in K^\sigma_c$ *and* $\neg\psi \in \mathrm{Cn}_0(\varphi)$ *then* $\neg \mathsf{E}_i \psi \in K^{\sigma \cdot \langle i, c, \psi \rangle}_c$.
2. *If* $\langle i, c, \varphi \rangle \in \sigma$ *and* $\mathsf{E}_i \varphi \in B^\sigma_c$ *then* $\varphi \in B^\sigma_c$.

Proof.

1. Suppose $\varphi \in K^\sigma_c$ and $\neg\psi \in \mathrm{Cn}_0(\varphi)$. Write $\rho = \sigma \cdot \langle i, c, \psi \rangle$. By **Soundness**, $\mathsf{S}_i \psi \in K^\rho_c$. By **K-conjunction**, $\varphi \in K^\sigma_c \subseteq (K^\sigma \sqcup K^{\langle i, c, \psi \rangle})_c \subseteq \mathrm{Cn}_c(K^\sigma \sqcup K^{\langle i, c, \psi \rangle}) = K^\rho_c$. Since $\neg\psi \in \mathrm{Cn}_0(\varphi)$ and $\varphi \in K^\rho_c$, **Closure** gives $\neg\psi \in K^\rho_c$. Recalling from Proposition 5.1.2 part (4) that $\mathsf{S}_i \psi \wedge \neg\psi \to \neg \mathsf{E}_i \psi$, **Closure** gives $\neg \mathsf{E}_i \psi \in K^\rho_c$, as desired.

2. Suppose $\langle i, c, \varphi \rangle \in \sigma$ and $\mathsf{E}_i \varphi \in B^\sigma_c$. By **Soundness** and **Containment**, $\mathsf{S}_i \varphi \in B^\sigma_c$. From Proposition 5.1.2 part (4) again we have $\mathsf{E}_i \varphi \wedge \mathsf{S}_i \varphi \to \varphi$. By **Closure**, $\varphi \in B^\sigma_c$.

□

(1) expresses how knowledge can negatively affect trust: we should distrust sources who make reports we know to be false. (2) expresses how trust affects belief: we should believe reports from trusted sources. It can also be seen as a form of *success* for ordinary sources, and implies AGM success when $i = *$ (by Proposition 5.1.2 part (6) and **Closure**). Note that these are the same ideas underlying the *coherence* axioms for truth discovery in Chapter 2. We illustrate the basic postulates by formalising the introductory hospital example.

Example 5.2.1. *Set* $\mathcal{S}, \mathcal{C}$ *and* Prop *as in Example 5.1.1, and consider the sequence*

$$\sigma = (\langle *, A, p \rangle, \langle \mathsf{X}, B, p \rangle, \langle \mathsf{Y}, B, \neg p \rangle, \langle \mathsf{X}, A, \neg p \rangle).$$

What do we know on the basis of this sequence, assuming the basic postulates? First note that by ***Soundness****, Proposition 5.1.2 part (6) and* ***Closure****, the report from* $*$ *gives* $p \in K^\sigma_A$*, i.e. reliable reports are known.* ***Soundness*** *also* *gives* $\mathsf{S_X}p \wedge \mathsf{S_Y}\neg p \in K^\sigma_B$. *Combined with Proposition 5.1.2 parts (2), (4) and* ***Closure****, this yields* $\neg(\mathsf{E_X}p \wedge \mathsf{E_Y}p) \in K^\sigma_c$ *for all* c*, formalising the intuitive idea that* X *and* Y *cannot both be experts on* p*, since they give conflicting reports. Considering the final report from* X*, we get* $p \wedge \mathsf{S_X}\neg p \in K^\sigma_A$*, and thus* $\neg\mathsf{E_X}p \in K^\sigma_c$ *by* ***Closure****. So in fact* X *is known to be a non-expert on* p*.*

What about beliefs? The basic postulates do not require beliefs to go beyond knowledge, so we cannot say much in general. An "optimistic" operator, however, may opt to believe that sources are experts unless we know otherwise, and thus maximise the information that can be (defeasibly) inferred from the sequence (in the next section we will introduce concrete operators obeying this principle). In this case we may believe that at least one source has expertise on p *(i.e.* $\mathsf{E_X}p \vee \mathsf{E_Y}p \in B^\sigma_c$*). Combined with* $\neg\mathsf{E_X}p \in K^\sigma_c$*,* ***Closure*** *and* ***Containment****, we get* $\mathsf{E_Y}p \in B^\sigma_B$*. Symmetry of expertise together with Proposition 5.2.3 part (2) then gives* $\neg p \in B^\sigma_B$*, i.e. we trust* Y *in the example and believe* B *does not suffer from condition* p*.*

5.2.2 Model-Based Operators

While an operator is a purely syntactic object, it will be convenient to specify K^σ and B^σ in semantic terms by selecting a set of *possible* and *most plausible* worlds for each sequence σ. We call such operators *model-based*.

Definition 5.2.2. *An operator is* model-based *if for every* σ *there are sets* $\mathcal{X}_\sigma, \mathcal{Y}_\sigma \subseteq \mathcal{W}$ *such that (i)* $\mathcal{X}_\sigma \supseteq \mathcal{Y}_\sigma$*; (ii)* $\Phi \in K^\sigma_c$ *iff* $W, c \models \Phi$ *for all* $W \in \mathcal{X}_\sigma$*; and (iii)* $\Phi \in B^\sigma_c$ *iff* $W, c \models \Phi$ *for all* $W \in \mathcal{Y}_\sigma$*.*

In other words, K^σ_c (resp., B^σ_c) contains the formulas which hold at case c in *all worlds* in $\mathcal{X}_\sigma$ (resp., $\mathcal{Y}_\sigma$). It follows from the relevant definitions that $\mathcal{X}_\sigma \subseteq \mathrm{mod}(K^\sigma)$, and equality holds if and only if $\mathcal{X}_\sigma$ is elementary (similarly for $\mathcal{Y}^\sigma$ and B^σ). Model-based operators are characterised by our first two basic postulates.

Theorem 5.2.1. *An operator satisfies* ***Closure*** *and* ***Containment*** *if and only if it is model-based.*

Proof. For ease of notation in what follows, write $\mathrm{mod}_c(\Phi) = \{W \in \mathcal{W} \mid W, c \models \Phi\}$.

"if": Suppose an operator $\sigma \mapsto \langle B^\sigma, K^\sigma\rangle$ is model-based. For **Closure**, we need to show that $B^\sigma_c \supseteq \mathrm{Cn}_c(B^\sigma)$ and $K^\sigma_c \supseteq \mathrm{Cn}_c(K^\sigma)$, for each c. Take any $\Phi \in \mathrm{Cn}_c(B^\sigma)$. Then $\mathrm{mod}(B^\sigma) \subseteq \mathrm{mod}_c(\Phi)$. From the relevant definitions, one can easily check that $\mathcal{Y}_\sigma \subseteq \mathrm{mod}(B^\sigma)$, so we have $\mathcal{Y}_\sigma \subseteq \mathrm{mod}_c(\Phi)$. That is, $W, c \models \Phi$ for all $W \in \mathcal{Y}_\sigma$. By definition of model-based operators, $\Phi \in B^\sigma_c$. The fact that $K^\sigma_c \supseteq \mathrm{Cn}_c(K^\sigma)$ follows by an identical argument upon noticing that $\mathcal{X}_\sigma \subseteq \mathrm{mod}(K^\sigma)$.

Containment follows from $\mathcal{X}_\sigma \supseteq \mathcal{Y}_\sigma$: if $\Phi \in K^\sigma_c$ then $W, c \models \Phi$ for all $W \in \mathcal{X}_\sigma$, and in particular this holds for all $W \in \mathcal{Y}_\sigma$. Hence $\Phi \in B^\sigma_c$, so $K^\sigma \sqsubseteq B^\sigma$.

"only if": Suppose an operator satisfies **Closure** and **Containment**. For any σ, set

$$\mathcal{X}_\sigma = \mathrm{mod}(K^\sigma)$$
$$\mathcal{Y}_\sigma = \mathrm{mod}(B^\sigma)$$

We show the three properties required in Definition 5.2.2. $\mathcal{X}_\sigma \supseteq \mathcal{Y}_\sigma$ follows from **Containment** and the definition of a model of a collection. For the second property, note that $\Phi \in K^\sigma_c$ iff $\Phi \in \mathrm{Cn}_c(K^\sigma)$ by **Closure**, i.e. iff $\mathrm{mod}(K^\sigma) \subseteq \mathrm{mod}_c(\Phi)$. By choice of $\mathcal{X}_\sigma$, this holds exactly when $W, c \models \Phi$ for all $W \in \mathcal{X}_\sigma$, as required. The third property is proved using an identical argument. □

Since we take **Closure** and **Containment** to be fundamental properties, all operators we consider from now on will be model-based. We introduce our first concrete operator.

Definition 5.2.3. *Define the model-based operator* $\mathsf{weak\text{-}mb}$ *by*

$$\mathcal{X}_\sigma = \mathcal{Y}_\sigma = \{W \mid W, c \models \mathsf{S}_i\varphi \text{ for all } \langle i, c, \varphi\rangle \in \sigma\}.$$

That is, the possible worlds $\mathcal{X}_\sigma$ are exactly those satisfying the soundness constraint for each report in σ, i.e. false reports are only due to lack of expertise of the corresponding source. Syntactically, we have $K^\sigma = B^\sigma = \mathrm{Cn}(G^\sigma_{\mathsf{snd}})$.

Proposition 5.2.4. $\mathsf{weak\text{-}mb}$ *satisfies all of the basic postulates.*

Proof (sketch). Most of the postulates are straightforward. Clearly **Closure** and **Containment** hold by Theorem 5.2.1. **Soundness** holds by construction. **K-bound** holds by monotonicity of Cn, since $K^\sigma = \mathrm{Cn}(G^\sigma_{\mathsf{snd}}) \sqsubseteq \mathrm{Cn}(G^\sigma_{\mathsf{snd}} \sqcup K^\emptyset)$. **Prior-extension** holds since $\mathcal{X}_\emptyset = \mathcal{W} \supseteq \mathcal{X}_\sigma$, so $K^\emptyset \sqsubseteq K^\sigma$. **Rearrangement** holds trivially, and **Equivalence** follows from the fact that $\mathsf{S}_i\varphi$ is equivalent to $\mathsf{S}_i\psi$ whenever $\varphi \equiv \psi$ (Proposition 5.1.2 (1)).

In fact, the only non-trivial postulate is **Consistency**. To show this holds, take any $*$-consistent sequence σ. By definition, for each case c the set of reports from $*$ is consistent, and is thus modelled by some valuation v_c. Forming a world W with these valuations and with each ordinary source $i \neq *$ having the trivial one-cell partition $\Pi_i = \{\mathcal{V}\}$, we have that *all* $\varphi \not\equiv \bot$ are sound for each i (since $\Pi_i[\varphi] = \mathcal{V} \ni v_c$), and all reports from $*$ are true, and in particular sound for $*$. Consequently $W \in \mathcal{X}_\sigma \subseteq \mathrm{mod}(K^\sigma)$, so $K^\sigma = B^\sigma$ are consistent. □

In fact, $\mathsf{weak\text{-}mb}$ is the *weakest* operator satisfying **Closure**, **Containment** and **Soundness**, in that for any other operator $\sigma \mapsto \langle \widehat{B}^\sigma, \widehat{K}^\sigma\rangle$ with these properties we have $B^\sigma \sqsubseteq \widehat{B}^\sigma$ and $K^\sigma \sqsubseteq \widehat{K}^\sigma$ for any σ. We come to an example.

Example 5.2.2. *Consider* weak-mb *applied to the sequence* $\sigma = (\langle *, c, p\rangle, \langle i, c, \neg p \wedge q\rangle)$. *By* ***Soundness***, ***Closure*** *and the validities from Proposition 5.1.2, we have* $p \in K_c^\sigma$ *and* $\neg \mathsf{E}_i p \in K_c^\sigma$*. In fact, by* ***Closure***, *we have* $\neg \mathsf{E}_i p \in K_d^\sigma$ *for all cases* d*. However, we cannot say much about* q*: neither* q, $\neg q$, $\mathsf{E}_i q$ *nor* $\neg \mathsf{E}_i q$ *are in* $B_c^\sigma = K_c^\sigma$.

5.3 Constructions

For model-based operators in Definition 5.2.2, the sets $\mathcal{X}_\sigma$ and $\mathcal{Y}_\sigma$ – which determine knowledge and belief – can depend on σ in a completely arbitrary manner. This lack of structure leads to very wide class of operators, and one cannot say much about them in general beyond the satisfaction of **Closure** and **Containment**. In this section we specialise model-based operators by providing two constructions.

5.3.1 Conditioning Operators

Intuitively, $\mathcal{Y}_\sigma$ is supposed to represent the *most plausible* worlds among the possible worlds in $\mathcal{X}_\sigma$. This suggests the presence of a *plausibility ordering* on $\mathcal{X}_\sigma$, which is used to select $\mathcal{Y}_\sigma$. For our first construction we take this approach: we condition a fixed plausibility total preorder on the knowledge $\mathcal{X}_\sigma$, and obtain $\mathcal{Y}_\sigma$ by selecting the minimal (i.e. most plausible) worlds.

Definition 5.3.1. *An operator is a* conditioning operator *if there is a total preorder* $\leq$ *on* $\mathcal{W}$ *and a mapping* $\sigma \mapsto \langle \mathcal{X}_\sigma, \mathcal{Y}_\sigma\rangle$ *as in Definition 5.2.2 such that* $\mathcal{Y}_\sigma = \min_{\leq} \mathcal{X}_\sigma$ *for all* σ.

Note that $\leq$ is independent of σ: it is fixed before receiving any reports. All conditioning operators are model-based by definition. Clearly $\mathcal{Y}_\sigma$ is determined by $\mathcal{X}_\sigma$ and the plausibility order, so that to define a conditioning operator it is enough to specify $\leq$ and the mapping $\sigma \mapsto \mathcal{X}_\sigma$. Write $W \simeq W'$ iff both $W \leq W'$ and $W' \leq W$.

Conditioning in this manner is well-established in the belief change literature. Our operators use a simplified case of conditionalisation as introduced by Spohn (1988). Boutilier, Friedman, and Halpern (1998) use a similar notion in their account of unreliable belief revision, wherein an agent's plausibility ranking is successively conditioned by its observations.

We now present examples of how the plausibility ordering $\leq$ may be defined.

Definition 5.3.2. *Define the conditioning operator* var-based-cond *by setting* $\mathcal{X}_\sigma$ *in the same way as* weak-mb *in Definition 5.2.3, and* $W \leq W'$ *iff* $r(W) \leq r(W')$*, where*

$$r(W) = -\sum_{i \in \mathcal{S}} \left|\left\{p \in \mathsf{Prop} \mid \Pi_i^W[p] = \|p\|\right\}\right|.$$

var-based-cond aims to trust each source on *as many propositional variables* as possible. One can check that var-based-cond satisfies the basic postulates, using essentially the same argument as for weak-mb in Proposition 5.2.4.

Example 5.3.1. *Revisiting the sequence $\sigma = (\langle *, c, p\rangle, \langle i, c, \neg p \wedge q\rangle)$ from Example 5.2.2 with* var-based-cond*, the knowledge set K_c^σ is the same as before, but we now have $q \wedge \mathsf{E}_i q \in B_c^\sigma$. This reflects the "credulous" behaviour of the ranking $\leq$: while it is not possible to believe i is an expert on p, we should believe they* are *an expert on q so long as this does not conflict with soundness. For the propositional beliefs generally, we have $[B_c^\sigma] = \mathrm{Cn}_0(p \wedge q)$. That is,* var-based-cond *takes the q part of the report from i (on which i is credulously trusted) while ignoring the $\neg p$ part (which is false due to report from $*$).*

Definition 5.3.3. *Define a conditioning operator* part-based-cond *with $\mathcal{X}_\sigma$ as for* var-based-cond*, and $\leq$ defined by the ranking function*

$$r(W) = -\sum_{i \in \mathcal{S}} |\Pi_i^W|.$$

part-based-cond aims to maximise the *number of cells* in the sources' partitions, and thereby maximise the number of propositions on which they have expertise. Unlike var-based-cond, the propositional variables play no special role. As expected, part-based-cond satisfies the basic postulates.

Example 5.3.2. *Applying* part-based-cond *to σ from Examples 5.2.2* and *5.3.1, we no longer extract q from the report of i: $q \notin B_c^\sigma$ and $\mathsf{E}_i q \notin B_c^\sigma$. Instead, we have $[B_c^\sigma] = \mathrm{Cn}_0(p)$, and $\mathsf{E}_i(p \vee q) \in B_c^\sigma$.*

Note that both var-based-cond and part-based-cond are based on the general principle of maximising the expertise of sources, subject to the constraint that all reports are sound. This intuition is formalised by the following postulate for conditioning operators. In what follows, write $W \preceq W'$ iff Π_i^W refines $\Pi_i^{W'}$ for all $i \in \mathcal{S}$, i.e. if all sources have broadly more expertise in W than in W'.[7]

Refinement. If $W \preceq W'$ then $W \leq W'$

Since $\preceq$ is only a partial order on $\mathcal{W}$ there are many possible total extensions; var-based-cond and part-based-cond provide two specific examples.

We now turn to an axiomatic characterisation of conditioning operators. Taken with the basic postulates from Section 5.2.1, conditioning operators can be characterised using an approach similar to that of Delgrande, Peppas, and Woltran (2018) in their account of *generalised AGM belief revision*.[8] This involves a technical property Delgrande, Peppas, and Woltran call **Acyc**, which finds its roots in the *Loop* property of Kraus, Lehmann, and Magidor (1990).

[7] Π refines Π' if $\forall A \in \Pi$, $\exists B \in \Pi'$ such that $A \subseteq B$.

[8] Note that while the result is similar, our framework is not an instance of theirs.

Inclusion-vacuity. $B^{\sigma \cdot \rho} \sqsubseteq \mathrm{Cn}(B^{\sigma} \sqcup K^{\rho})$, with equality if $B^{\sigma} \sqcup K^{\rho}$ is consistent

Acyc. If $\sigma_0, \ldots, \sigma_n$ are such that $K^{\sigma_j} \sqcup B^{\sigma_{j+1}}$ is consistent for all $0 \leq j < n$ and $K^{\sigma_n} \sqcup B^{\sigma_0}$ is consistent, then $K^{\sigma_0} \sqcup B^{\sigma_n}$ is consistent

Inclusion-vacuity is so-named since it is analogous to the combination of *Inclusion* and *Vacuity* from AGM revision, if one informally views $B^{\sigma \cdot \rho}$ as the revision of B^{σ} by K^{ρ}. **Acyc** is the analogue of the postulate of Delgrande, Peppas, and Woltran, which rules out cycles in the plausibility order constructed in the representation result.

As with the result of Delgrande, Peppas, and Woltran, a technical condition beyond Definition 5.3.1 is required to obtain the characterisation: say that a conditioning operator is *elementary* if for each σ the sets of worlds $\mathcal{X}_{\sigma}$ and $\mathcal{Y}_{\sigma} = \min_{\leq} \mathcal{X}_{\sigma}$ are elementary.[9]

Theorem 5.3.1. *Suppose an operator satisfies the basic postulates of Section 5.2.1. Then it is an elementary conditioning operator if and only if it satisfies **Inclusion-vacuity** and **Acyc**.*

The proof is roughly follows the lines of Theorem 4.9 in (Delgrande, Peppas, and Woltran 2018), although some differences arise due to the form of our input as finite sequences of reports. In fact, we will prove a result more general than Theorem 5.3.1, which does not require the full set of basic postulates to hold. For example, the general result applies to conditioning operators which do not necessarily satisfy **Soundness**. In order to state the result in full generality, we introduce two auxiliary postulates – each of which follows from the basic postulates and **Inclusion-vacuity**.

Conditional-consistency. If K^{σ} is consistent then so is B^{σ}

Duplicate-removal. If $\langle i, c, \varphi \rangle \in \sigma$ then $B^{\sigma \cdot \langle i, c, \varphi \rangle} = B^{\sigma}$ and $K^{\sigma \cdot \langle i, c, \varphi \rangle} = K^{\sigma}$

Conditional-consistency is another consistency postulate which says beliefs B^{σ} can only be inconsistent if the knowledge K^{σ} is too. **Duplicate-removal** says that the output does not change when a report $\langle i, c, \varphi \rangle$ in σ is repeated; together with **Rearrangement**, this means any duplicate reports may be ignored.

Lemma 5.3.1. *Any operator satisfying the basic postulates and **Inclusion-vacuity** also satisfies both **Conditional-consistency** and **Duplicate-removal**.*

Proof. For **Conditional-consistency** we show the contrapositive. Suppose B^{σ} is inconsistent. By **Consistency**, σ cannot be $*$-consistent. Thus, there is

[9] Equivalently, there is a total preorder $\leq$ such that $\mathrm{mod}(B^{\sigma}) = \min_{\leq} \mathrm{mod}(K^{\sigma})$ for all σ.

some case c such that the reports from $*$ are inconsistent. But by **Soundness** and **Closure**, $\langle *, c, \varphi\rangle \in \sigma$ implies $\varphi \in K^\sigma_c$. Hence K^σ is inconsistent.

For **Duplicate-removal**, suppose $\langle i, c, \varphi\rangle \in \sigma$. First note that if σ is $*$-inconsistent, the argument above shows K^σ is inconsistent. By **Containment**, so too is B^σ. By **Closure**, both K^σ and B^σ contain *all* formulas of $\mathcal{L}$. But $\sigma \cdot \langle i, c, \varphi\rangle$ is also $*$-inconsistent, so $B^{\sigma\cdot\langle i,c,\varphi\rangle} = B^\sigma$ and $K^{\sigma\cdot\langle i,c,\varphi\rangle} = K^\sigma$.

So, suppose σ is $*$-consistent. By **Consistency**, B^σ is consistent. By **Rearrangement** and **K-conjunction**, $\langle i, c, \varphi\rangle \in \sigma$ implies $K^{\langle i,c,\varphi\rangle} \sqsubseteq K^\sigma$. Appealing to **Containment**, $K^{\langle i,c,\varphi\rangle} \sqsubseteq B^\sigma$. Thus

$$\mathrm{Cn}(B^\sigma \sqcup K^{\langle i,c,\varphi\rangle}) = \mathrm{Cn}(B^\sigma) = B^\sigma$$

with **Closure** applied in the final step. Since this is a consistent collection, **Inclusion-vacuity** yields

$$B^{\sigma\cdot\langle i,c,\varphi\rangle} = \mathrm{Cn}(B^\sigma \sqcup K^{\langle i,c,\varphi\rangle}) = B^\sigma$$

as required. For knowledge, pairing $K^{\langle i,c,\varphi\rangle} \sqsubseteq K^\sigma$ with **K-conjunction** gives

$$K^{\sigma\cdot\langle i,c,\varphi\rangle} = \mathrm{Cn}(K^\sigma \sqcup K^{\langle i,c,\varphi\rangle}) = \mathrm{Cn}(K^\sigma) = K^\sigma,$$

which completes the proof. □

Lemma 5.3.2. *For any model-based operator and sequence σ, $\mathcal{X}_\sigma = \mathrm{mod}(K^\sigma)$ iff $\mathcal{X}_\sigma$ is elementary, and $\mathcal{Y}_\sigma = \mathrm{mod}(B^\sigma)$ iff $\mathcal{Y}_\sigma$ is elementary.*

Proof. We prove the result for $\mathcal{X}_\sigma$ and K^σ only. The "only if" direction is clear from the definition of an elementary set. For the "if" direction, suppose $\mathcal{X}_\sigma$ is elementary, i.e. $\mathcal{X}_\sigma = \mathrm{mod}(G)$ for some collection G. Since $\Phi \in K^\sigma_c$ iff $\mathcal{X}_\sigma \subseteq \mathrm{mod}_c(\Phi)$, we have $K^\sigma_c = \mathrm{Cn}_c(G)$, i.e. $K^\sigma = \mathrm{Cn}(G)$. Consequently $\mathrm{mod}(K^\sigma) = \mathrm{mod}(\mathrm{Cn}(G)) = \mathrm{mod}(G) = \mathcal{X}_\sigma$. □

We now come to the more general characterisation. Once proved, the main result in Theorem 5.3.1 easily follows in light of Proposition 5.2.2 and Lemma 5.3.1.

Proposition 5.3.1. *Suppose an operator satisfies **Closure**, **Containment**, **K-conjunction** and **Equivalence**. Then it is an elementary conditioning operator if and only if it satisfies **Rearrangement**, **Duplicate-removal**, **Conditional-consistency**, **Inclusion-vacuity** and **Acyc**.*

Proof. Take some operator $\sigma \mapsto \langle B^\sigma, K^\sigma\rangle$ satisfying **Closure**, **Containment**, **K-conjunction** and **Equivalence**.

"if": Suppose the operator in question additionally satisfies **Rearrangement**, **Duplicate-removal**, **Conditional-consistency**, **Inclusion-vacuity** and **Acyc**. For any σ, set

$$\mathcal{X}_\sigma = \mathrm{mod}(K^\sigma)$$
$$\mathcal{Y}_\sigma = \mathrm{mod}(B^\sigma)$$

Then – by **Closure** and **Containment** as shown in the proof of Theorem 5.2.1 – our operator is model based corresponding to this choice of $\mathcal{X}_\sigma$ and $\mathcal{Y}_\sigma$. Clearly both are elementary. We will construct a total preorder $\leq$ over $\mathcal{W}$ such that $\mathcal{Y}_\sigma = \min_\leq \mathcal{X}_\sigma$; this will show the operator is an elementary conditioning operator.

First, fix a function $c : \mathcal{L}_0/{\equiv} \to \mathcal{L}_0$ which chooses a fixed representative of each equivalence class of logically equivalent propositional formulas, i.e. any mapping such that $c([\varphi]_\equiv) \equiv \varphi$. To simplify notation, write $\widehat{\varphi}$ for $c([\varphi]_\equiv)$. Then $\varphi \equiv \widehat{\varphi}$. Write $\widehat{\mathcal{L}_0} = \{\widehat{\varphi} \mid \varphi \in \mathcal{L}_0\}$. Note that $\widehat{\mathcal{L}_0}$ is finite (since we work with only finitely many propositional variables) and every formula in $\mathcal{L}_0$ is equivalent to exactly one formula in $\widehat{\mathcal{L}_0}$. For a sequence σ, let $\widehat{\sigma}$ be the result of replacing each report $\langle i, c, \varphi\rangle$ with $\langle i, c, \widehat{\varphi}\rangle$. Note that by **Rearrangement** and **Equivalence**, $\mathcal{X}_{\widehat{\sigma}} = \mathcal{X}_\sigma$ and $\mathcal{Y}_{\widehat{\sigma}} = \mathcal{Y}_\sigma$.

Now, for any world W, set

$$\mathcal{R}(W) = \{\langle i, c, \varphi\rangle \in \mathcal{S} \times \mathcal{C} \times \widehat{\mathcal{L}_0} \mid W \in \mathcal{X}_{\langle i,c,\varphi\rangle}\}$$

Note that $\mathcal{R}(W)$ is finite. For any pair of worlds W_1, W_2, let $\rho(W_1, W_2)$ be some enumeration of $\mathcal{R}(W_1) \cap \mathcal{R}(W_2)$. We establish some useful properties of $\rho(W_1, W_2)$.

Claim 5.3.1. *If $\rho(W_1, W_2) \neq \emptyset$, $W_1, W_2 \in \mathcal{X}_{\rho(W_1,W_2)}$.*

Proof of claim. By **K-conjunction**, for any sequences σ, ρ we have $K^{\sigma\cdot\rho} = \mathrm{Cn}(K^\sigma \sqcup K^\rho)$. Taking the models of both sides, we have $\mathcal{X}_{\sigma\cdot\rho} = \mathcal{X}_\sigma \cap \mathcal{X}_\rho$. It follows that for $\rho(W_1, W_2) \neq \emptyset$,

$$\mathcal{X}_{\rho(W_1,W_2)} = \bigcap_{\langle i,c,\varphi\rangle \in \rho(W_1,W_2)} \mathcal{X}_{\langle i,c,\varphi\rangle}$$

If $\langle i, c, \varphi\rangle \in \rho(W_1, W_2)$ then $W_1, W_2 \in \mathcal{X}_{\langle i,c,\varphi\rangle}$ by definition. Hence $W_1, W_2 \in \mathcal{X}_{\rho(W_1,W_2)}$. ◇

Claim 5.3.2. *If a sequence σ contains no equivalent reports (i.e. no distinct tuples $\langle i, c, \varphi\rangle$, $\langle i, c, \psi\rangle$ with $\varphi \equiv \psi$) and $W_1, W_2 \in \mathcal{X}_\sigma$, there is a sequence δ such that $W_1, W_2 \in \mathcal{X}_\delta$ and $\rho(W_1, W_2)$ is a permutation of $\widehat{\sigma} \cdot \delta$.*

Proof of claim. If $\sigma = \emptyset$ then we can simply take $\delta = \rho(W_1, W_2)$. So suppose $\sigma \neq \emptyset$. By the same argument as in the proof of Claim 5.3.1, we have

$$\mathcal{X}_\sigma = \bigcap_{\langle i,c,\varphi\rangle \in \sigma} \mathcal{X}_{\langle i,c,\varphi\rangle}$$

Take any $\langle i, c, \varphi\rangle \in \widehat{\sigma}$. Then $\varphi \in \widehat{\mathcal{L}_0}$, and there is $\psi \equiv \varphi$ such that $\langle i, c, \psi\rangle \in \sigma$. By **Equivalence**, we have

$$W_1, W_2 \in \mathcal{X}_\sigma \subseteq \mathcal{X}_{\langle i,c,\psi\rangle} = \mathcal{X}_{\langle i,c,\varphi\rangle}$$

i.e. $\langle i, c, \varphi \rangle \in \mathcal{R}(W_1) \cap \mathcal{R}(W_2)$. Hence $\langle i, c, \varphi \rangle$ appears in $\rho(W_1, W_2)$. By the assumption that σ contains no equivalent reports, $\widehat{\sigma}$ contains no duplicates. It follows that $\rho(W_1, W_2)$ can be permuted so that $\widehat{\sigma}$ appears as a prefix. Taking δ to be the sequence that remains after $\widehat{\sigma}$ in this permutation, we clearly have that $\rho(W_1, W_2)$ is a permutation of $\widehat{\sigma} \cdot \delta$. Since $\sigma \neq \emptyset$ implies $\widehat{\sigma} \neq \emptyset$ and thus $\rho(W_1, W_2) \neq \emptyset$, by **Rearrangement**, **K-conjunction** and Claim 5.3.1 we get

$$W_1, W_2 \in \mathcal{X}_{\rho(W_1,W_2)} = \mathcal{X}_{\widehat{\sigma}\cdot\delta} = \mathcal{X}_{\widehat{\sigma}} \cap \mathcal{X}_\delta \subseteq \mathcal{X}_\delta$$

and we are done. $\diamondsuit$

Now define a relation R on $\mathcal{W}$ by

$$WRW' \iff W = W' \text{ or } W \in \mathcal{Y}_{\rho(W,W')}$$

We have that any world in $\mathcal{Y}_\sigma$ R-precedes all worlds $\mathcal{X}_\sigma$.

Claim 5.3.3. *If $W \in \mathcal{Y}_\sigma$, then for all $W' \in \mathcal{X}_\sigma$ we have WRW'*

Proof of claim. By **Rearrangement**, **Equivalence** and **Duplicate-removal**, we may assume without loss of generality that σ contains no distinct equivalent reports.

Let $W \in \mathcal{Y}_\sigma$ and $W' \in \mathcal{X}_\sigma$. Then $W \in \mathcal{X}_\sigma$ too. By Claim 5.3.2 and **Rearrangement**, there is some sequence δ such that $\mathcal{Y}_{\rho(W,W')} = \mathcal{Y}_{\widehat{\sigma}\cdot\delta}$ and $W, W' \in \mathcal{X}_\delta$. Consequently $W \in \mathcal{Y}_\sigma \cap \mathcal{X}_\delta = \mathcal{Y}_{\widehat{\sigma}} \cap \mathcal{X}_\delta$. Thus $B^{\widehat{\sigma}} \sqcup K^\delta$ is consistent. From **Inclusion-vacuity** we get

$$\mathcal{Y}_{\widehat{\sigma}\cdot\delta} = \mathcal{Y}_{\widehat{\sigma}} \cap \mathcal{X}_\delta$$

Thus

$$W \in \mathcal{Y}_{\widehat{\sigma}} \cap \mathcal{X}_\delta = \mathcal{Y}_{\widehat{\sigma}\cdot\delta} = \mathcal{Y}_{\rho(W,W')}$$

so WRW' as required. $\diamondsuit$

Now let $\leq_0$ be the transitive closure of R. Then $\leq_0$ is a (partial) preorder. By Claim 5.3.3, every world in $\mathcal{Y}_\sigma$ is $\leq_0$-minimal in $\mathcal{X}_\sigma$. In fact, the converse is also true.

Claim 5.3.4. *If $W \in \mathcal{X}_\sigma$ and there is no $W' \in \mathcal{X}_\sigma$ with $W' <_0 W$, then $W \in \mathcal{Y}_\sigma$.*

Proof of claim. As before, assume without loss of generality that σ contains no distinct equivalent reports.

Take W as in the statement of the claim. Then $\mathcal{X}_\sigma \neq \emptyset$, so $\mathcal{Y}_\sigma \neq \emptyset$ by **Conditional-consistency**. Let $W' \in \mathcal{Y}_\sigma$. By Claim 5.3.3, $W'RW$ and thus $W' \leq_0 W$. But by assumption, $W' \not<_0 W$. So we must have $W \leq_0 W'$. By definition of $\leq_0$ as the transitive closure of R, there are $W_0, \ldots W_n$ such that $W_0 = W$, $W_n = W'$ and

$$W_j R W_{j+1} \qquad (0 \leq j < n)$$

Without loss of generality, $n > 0$ and each of the W_j are distinct. From the definition of R, we therefore have that

$$W_j \in \rho(W_j, W_{j+1}) \qquad (0 \leq j < n)$$

Now set

$$\begin{aligned} \rho_j &= \rho(W_j, W_{j+1}) \qquad (0 \leq j < n) \\ \rho_n &= \rho(W_0, W_n) \end{aligned}$$

Since $W'RW$, i.e. $W_n R W_0$, we in fact have $W_j \in \mathcal{Y}_{\rho_j}$ for all j (including $j = n$). For $j < n$, we also have $W_{j+1} \in \mathcal{X}_{\rho_j}$.[10] Consequently, for $j < n$ we have

$$W_{j+1} \in \mathcal{X}_{\rho_j} \cap \mathcal{Y}_{\rho_{j+1}}$$

i.e. $K^{\rho_j} \sqcup B^{\rho_{j+1}}$ is consistent. Moreover, $W_0 \in \mathcal{X}_{\rho_n} \cap \mathcal{Y}_{\rho_0}$, so $K^{\rho_n} \sqcup B^{\rho_0}$ is consistent. We can now apply **Acyc**: we get that $K^{\rho_0} \sqcup B^{\rho_n}$ is also consistent. On the one hand, **Inclusion-vacuity** and consistency of $K^{\rho_n} \sqcup B^{\rho_0}$ gives

$$B^{\rho_0 \cdot \rho_n} = \mathrm{Cn}(B^{\rho_0} \sqcup K^{\rho_n})$$

On the other, consistency of $B^{\rho_n} \sqcup K^{\rho_0}$ and **Rearrangement** gives

$$B^{\rho_0 \cdot \rho_n} = B^{\rho_n \cdot \rho_0} = \mathrm{Cn}(B^{\rho_n} \sqcup K^{\rho_0})$$

Combining these and taking models, we find

$$\mathcal{Y}_{\rho_0} \cap \mathcal{X}_{\rho_n} = \mathcal{Y}_{\rho_n} \cap \mathcal{X}_{\rho_0}$$

In particular, since W_0 lies in the set on the left-hand side, we have $W_0 \in \mathcal{Y}_{\rho_n}$.

Now, since $W_0, W_n \in \mathcal{X}_\sigma$ and $\rho_n = \rho(W_0, W_n)$, Claim 5.3.2 gives that there is δ with $W_0, W_n \in \mathcal{X}_\delta$ such that ρ_n is a permutation of $\widehat{\sigma} \cdot \delta$. Recalling that $W_n = W' \in \mathcal{Y}_\sigma = \mathcal{Y}_{\widehat{\sigma}}$ by assumption, we have $W_n \in \mathcal{Y}_{\widehat{\sigma}} \cap \mathcal{X}_\delta$, i.e. $B^{\widehat{\sigma}} \sqcup K^\delta$ is consistent. Applying **Inclusion-vacuity** once more, we get

$$B^{\rho_n} = B^{\widehat{\sigma} \cdot \delta} = \mathrm{Cn}(B^{\widehat{\sigma}} \sqcup K^\delta) = \mathrm{Cn}(B^\sigma \sqcup K^\delta)$$

Taking models of both sides,

$$\mathcal{Y}_{\rho_n} = \mathcal{Y}_\sigma \cap \mathcal{X}_\delta \subseteq \mathcal{Y}_\sigma$$

But we already saw that $W_0 \in \mathcal{Y}_{\rho_n}$. Hence $W_0 \in \mathcal{Y}_\sigma$. Since $W_0 = W$, we are done. $\diamondsuit$

[10] If $\rho_j \neq \emptyset$ this follows from Claim 5.3.1. Otherwise, $W_{j+1} \in \mathcal{Y}_{\rho_{j+1}} \subseteq \mathcal{X}_{\rho_{j+1}} = \mathcal{X}_{\rho_{j+1} \cdot \emptyset} = \mathcal{X}_{\rho_{j+1}} \cap \mathcal{X}_\emptyset \subseteq \mathcal{X}_\emptyset = \mathcal{X}_{\rho_j}$ by **K-conjunction**.

To complete the proof we extend $\leq_0$ to a *total* preorder and show that this does not affect the minimal elements of each $\mathcal{X}_\sigma$. Indeed, let $\leq$ be any total preorder extending $\leq_0$ and preserving strict inequalities, i.e. $\leq$ such that (i) $W \leq_0 W'$ implies $W \leq W'$; and (ii) $W <_0 W'$ implies $W < W'$.[11]

Claim 5.3.5. *For any sequence σ, $\mathcal{Y}_\sigma = \min_{\leq} \mathcal{X}_\sigma$*

Proof of claim. Take any σ. For the left-to-right inclusion, take $W \in \mathcal{Y}_\sigma$. Then $W \in \mathcal{X}_\sigma$. Let $W' \in \mathcal{X}_\sigma$. By Claim 5.3.3, WRW', so $W \leq_0 W'$ and $W \leq W'$. Hence W is $\leq$-minimal in $\mathcal{X}_\sigma$.

For the right-to-left inclusion, take $W \in \min_{\leq} \mathcal{X}_\sigma$. Then for any $W' \in \mathcal{X}_\sigma$ we have $W \leq W'$. In particular, $W' \not< W$. By property (ii) of $\leq$, we have $W' \not<_0 W$. Since W' was an arbitrary member of $\mathcal{X}_\sigma$ and $W \in \mathcal{X}_\sigma$, the conditions of Claim 5.3.4 are satisfied, and we get $W \in \mathcal{Y}_\sigma$. $\diamondsuit$

This shows that our operator is an elementary conditioning operator as required.

"only if": Now suppose the operator is an elementary conditioning operator. i.e. there is a total preorder $\leq$ on $\mathcal{W}$ and a mapping $\sigma \mapsto \langle \mathcal{X}_\sigma, \mathcal{Y}_\sigma \rangle$ such that for each σ, $\mathcal{Y}_\sigma = \min_{\leq} X_\sigma$, $\mathcal{X}_\sigma$ and $\mathcal{Y}_\sigma$ are elementary, and K^σ, B^σ are determined by $\mathcal{X}_\sigma$, $\mathcal{Y}_\sigma$ respectively according to Definition 5.2.2. By elementariness and Lemma 5.3.2, $\mathcal{X}_\sigma = \mathrm{mod}(K^\sigma)$ and $\mathcal{Y}_\sigma = \mathrm{mod}(B^\sigma)$.

The following claim will be useful at various points.

Claim 5.3.6. *Suppose σ and ρ are such that $\mathcal{X}_\sigma = \mathcal{X}_\rho$. Then $K^\sigma = K^\rho$ and $B^\sigma = B^\rho$.*

Proof of claim. Since the total preorder $\leq$ is fixed, we have

$$Y_\sigma = \min_{\leq} \mathcal{X}_\sigma = \min_{\leq} \mathcal{X}_\rho = \mathcal{Y}_\rho$$

Now, $\mathcal{X}_\sigma = \mathcal{X}_\rho$ means $\mathrm{mod}(K^\sigma) = \mathrm{mod}(K^\rho)$, so $\mathrm{Cn}(K^\sigma) = \mathrm{Cn}(K^\rho)$. By **Closure**, $K^\sigma = K^\rho$. Similarly, $\mathcal{Y}_\sigma = \mathcal{Y}_\rho$ gives $B^\sigma = B^\rho$. $\diamondsuit$

We take the postulates to be shown in turn.

- **Rearrangement**: Suppose σ is a permutation of ρ. Without loss of generality, $\sigma, \rho \neq \emptyset$. Repeated application of **K-conjunction** gives

  $$\mathcal{X}_\sigma = \bigcap_{\langle i,c,\varphi \rangle \in \sigma} \mathcal{X}_{\langle i,c,\varphi \rangle}$$

  Since σ and ρ contain exactly the same reports – just in a different order – commutativity and associativity of intersection of sets gives $\mathcal{X}_\sigma = \mathcal{X}_\rho$. **Rearrangement** follows from Claim 5.3.6.

[11] Such $\leq$ always exists. Indeed, note that $\leq_0$ induces a partial order on the equivalence classes of $\mathcal{W}$ with respect to the symmetric part of $\leq_0$ given by $W \simeq_0 W'$ iff $W \leq_0 W'$ and $W' \leq_0 W$. This partial order can be extended to a linear order $\leq^*$ on the equivalence classes. Taking $W \leq W'$ iff $[W] \leq^* [W']$, we obtain a total preorder on $\mathcal{W}$ with the desired properties.

- **Duplicate-removal**: Suppose $\langle i, c, \varphi \rangle \in \sigma$. Then there is a (possibly empty) sequence ρ such that σ is a permutation of $\rho \cdot \langle i, c, \varphi \rangle$. By **Rearrangement** just shown and **K-conjunction**, we have

$$\begin{aligned}
\mathcal{X}_{\sigma \cdot \langle i,c,\varphi \rangle} &= \mathcal{X}_{\rho \cdot \langle i,c,\varphi \rangle \cdot \langle i,c,\varphi \rangle} \\
&= \mathcal{X}_{\rho} \cap \mathcal{X}_{\langle i,c,\varphi \rangle} \cap \mathcal{X}_{\langle i,c,\varphi \rangle} \\
&= \mathcal{X}_{\rho} \cap \mathcal{X}_{\langle i,c,\varphi \rangle} \\
&= \mathcal{X}_{\rho \cdot \langle i,c,\varphi \rangle} \\
&= \mathcal{X}_{\sigma}
\end{aligned}$$

  and we may conclude by Claim 5.3.6.

- **Conditional-consistency**: Suppose K^{σ} is consistent, i.e. $\mathcal{X}_{\sigma} \neq \emptyset$. Since $\mathcal{W}$ is finite, $\mathcal{X}_{\sigma}$ is finite and thus some $\leq$-minimal world must exist in $\mathcal{X}_{\sigma}$. Hence $\mathcal{Y}_{\sigma} \neq \emptyset$, so B^{σ} is consistent.

- **Inclusion-vacuity**: Take any sequences σ, ρ. First we show $B^{\sigma \cdot \rho} \sqsubseteq \mathrm{Cn}(B^{\sigma} \sqcup K^{\rho})$, or equivalently, $\mathcal{Y}_{\sigma \cdot \rho} \supseteq \mathcal{Y}_{\sigma} \cap \mathcal{X}_{\rho}$. Suppose $W \in \mathcal{Y}_{\sigma} \cap \mathcal{X}_{\rho}$. Since $\mathcal{Y}_{\sigma} \subseteq \mathcal{X}_{\sigma}$, we have $W \in \mathcal{X}_{\sigma} \cap \mathcal{X}_{\rho} = \mathcal{X}_{\sigma \cdot \rho}$ by **K-conjunction**. We need to show W is minimal. Take any $W' \in \mathcal{X}_{\sigma \cdot \rho}$. Then $W' \in \mathcal{X}_{\sigma}$, so $W \in \mathcal{Y}_{\sigma} = \min_{\leq} \mathcal{X}_{\sigma}$ gives $W \leq W'$. Hence $W \in \min_{\leq} X_{\sigma \cdot \rho} = \mathcal{Y}_{\sigma \cdot \rho}$.

  Now suppose $B^{\sigma} \sqcup K^{\rho}$ is consistent, i.e. $\mathcal{Y}_{\sigma} \cap \mathcal{X}_{\rho} \neq \emptyset$. Take some $\widehat{W} \in \mathcal{Y}_{\sigma} \cap \mathcal{X}_{\rho}$. We need to show $B^{\sigma \cdot \rho} \sqsupseteq \mathrm{Cn}(B^{\sigma} \sqcup K^{\rho})$, i.e. $\mathcal{Y}_{\sigma \cdot \rho} \subseteq \mathcal{Y}_{\sigma} \cap \mathcal{X}_{\rho}$. To that end, let $W \in \mathcal{Y}_{\sigma \cdot \rho}$. Then $W \in \mathcal{X}_{\sigma \cdot \rho} = \mathcal{X}_{\sigma} \cap \mathcal{X}_{\rho} \subseteq \mathcal{X}_{\rho}$, so we only need to show $W \in \mathcal{Y}_{\sigma}$. Take any $W' \in \mathcal{X}_{\sigma}$. Then $\widehat{W} \in \mathcal{Y}_{\sigma}$ gives $\widehat{W} \leq W'$. But $\widehat{W} \in \mathcal{X}_{\sigma} \cap \mathcal{X}_{\rho} = \mathcal{X}_{\sigma \cdot \rho}$ and $W \in \mathcal{Y}_{\sigma \cdot \rho}$ gives $W \leq \widehat{W}$. By transitivity of $\leq$, we have $W \leq W'$. Hence $W \in \min_{\leq} \mathcal{X}_{\sigma} = \mathcal{Y}_{\sigma}$.

- **Acyc**: Let $\sigma_0, \ldots, \sigma_n$ be as in the statement of **Acyc**. Without loss of generality, $n > 0$. Then there are $W_0, \ldots, W_n$ such that

$$\begin{aligned}
&W_j \in \mathcal{X}_{\sigma_j} \cap \mathcal{Y}_{\sigma_{j+1}} \qquad (0 \leq j < n) \\
&W_n \in \mathcal{X}_{\sigma_n} \cap \mathcal{Y}_{\sigma_0}
\end{aligned}$$

  Note that $W_j \in \mathcal{X}_{\sigma_j}$ for all j. For $j < n$, we also have $W_j \in \mathcal{Y}_{\sigma_{j+1}} = \min_{\leq} X_{\sigma_{j+1}}$. It follows that $W_j \leq W_{j+1}$ for such j, so

$$W_0 \leq \cdots \leq W_n$$

  But we also have $W_n \in \mathcal{Y}_{\sigma_0} = \min_{\leq} \mathcal{X}_{\sigma_0}$ and $W_0 \in \mathcal{X}_{\sigma_0}$, so $W_n \leq W_0$. By transitivity of $\leq$, the chain flattens: we have

$$W_0 \simeq \cdots \simeq W_n$$

Now note that since $W_{n-1} \in \mathcal{Y}_{\sigma_n}$, W_{n-1} is minimal in $\mathcal{X}_{\sigma_n}$. But $W_n \in \mathcal{X}_{\sigma_n}$ and $W_{n-1} \simeq W_n$ by the above, so in fact $W_n \in \mathcal{Y}_{\sigma_n}$ too. Hence

$$\begin{aligned} W_n &\in \mathcal{Y}_{\sigma_0} \cap \mathcal{Y}_{\sigma_n} \\ &\subseteq \mathcal{X}_{\sigma_0} \cap \mathcal{Y}_{\sigma_n} \\ &= \mathrm{mod}(K^{\sigma_0} \sqcup B^{\sigma_n}) \end{aligned}$$

i.e. $K^{\sigma_0} \sqcup B^{\sigma_n}$ is consistent, as required for **Acyc**.

□

Note that while the requirement in Theorem 5.3.1 that $\mathcal{X}_\sigma$ and $\mathcal{Y}_\sigma$ are elementary is a technical condition,[12] the characterisation in Proposition 5.1.4 implies a simple sufficient condition for elementariness.

Proposition 5.3.2. *Suppose $\leq$ is such that $W \simeq W'$ whenever W and W' are partition-equivalent. Then $\min_\leq S$ is elementary for any elementary set $S \subseteq \mathcal{W}$.*

Proof. We use the characterisation of elementary sets from Proposition 5.1.4. Take $S \subseteq \mathcal{W}$ elementary. Suppose $W \in \mathcal{W}$, $W_1, W_2 \in \min_\leq S$ are such that W is partition-equivalent to both W_1, W_2 and W is a valuation combination from $\{W_1, W_2\}$. By hypothesis we have $W \simeq W_1 \simeq W_2$.

Now since $\min_\leq S \subseteq S$, we have $W_1, W_2 \in S$. Since S is elementary, $W \in S$. But now $W \simeq W_1$ and $W_1 \in \min_\leq S$ gives $W \in \min_\leq S$. This shows the required closure property for $\min_\leq S$, and we are done. □

Proposition 5.3.2 implies that $\mathsf{var\text{-}based\text{-}cond}$ and $\mathsf{part\text{-}based\text{-}cond}$ are elementary. Indeed, for both operators $\mathcal{X}_\sigma = \mathrm{mod}(G^\sigma_{\mathsf{snd}})$ so is elementary by definition. Since the ranking $\leq$ for each operator only depends on the partitions of worlds, $\mathcal{Y}_\sigma = \min_\leq \mathcal{X}_\sigma$ is elementary also.

5.3.2 Score-Based Operators

The fact that the plausibility order $\leq$ of a conditioning operator is fixed may be too limiting. For example, consider

$$\sigma = (\langle i, c, p \rangle, \langle j, c, \neg p \rangle, \langle i, d, p \rangle).$$

If one sets $\mathcal{X}_\sigma$ to satisfy the soundness constraints (i.e. as in $\mathsf{weak\text{-}mb}$), there is a possible world $W_1 \in \mathcal{X}_\sigma$ with $W_1, d \models \neg \mathsf{E}_i p \wedge \mathsf{E}_j p \wedge \neg p$ (i.e. W_1 sides with source j and p is false at d) and another world $W_2 \in \mathcal{X}_\sigma$ with $W_2, d \models \mathsf{E}_i p \wedge \neg \mathsf{E}_j p \wedge p$ (i.e. W_2 sides with source i). Appealing to symmetry, one may argue that neither world is *a priori* more plausible than the other, so any fixed plausibility order should have $W_1 \simeq W_2$. If these worlds are maximally plausible (e.g. if taking

[12] **Inclusion-vacuity** may fail for non-elementary conditioning.

the "optimistic" view outlined in Example 5.2.1), conditioning gives $p \notin B_d^\sigma$ and $\neg p \notin B_d^\sigma$. However, there is an argument that W_2 should be considered more plausible than W_1 *given the sequence* σ, since W_2 validates the final report $\langle i, d, p\rangle$ whereas W_1 does not. Consequently, there is an argument that we should in fact have $p \in B_d^\sigma$.[13] This shows that we need the plausibility order to be responsive to the input sequence for adequate belief change.[14]

As a result of this discussion, we look for operators whose plausibility ordering can depend on σ. One approach to achieve this in a controlled way is to have a ranking for each *report* $\langle i, c, \varphi\rangle$, and combine these to construct a ranking for each sequence σ. We represent these rankings by *scoring functions*, and call the resulting operators *score-based*.

Definition 5.3.4. *An operator is* score-based *if there is a mapping* $\sigma \mapsto \langle \mathcal{X}_\sigma, \mathcal{Y}_\sigma\rangle$ *as in Definition 5.2.2 and functions* $r_0 : \mathcal{W} \to \mathbb{N}_0 \cup \{\infty\}$, $d : \mathcal{W} \times (\mathcal{S} \times \mathcal{C} \times \mathcal{L}_0) \to \mathbb{N}_0 \cup \{\infty\}$ *such that* $\mathcal{X}_\sigma = \{W \mid r_\sigma(W) < \infty\}$ *and* $\mathcal{Y}_\sigma = \operatorname{argmin}_{W \in \mathcal{X}_\sigma} r_\sigma(W)$, *where*

$$r_\sigma(W) = r_0(W) + \sum\nolimits_{\langle i,c,\varphi\rangle \in \sigma} d(W, \langle i, c, \varphi\rangle).$$

Here $r_0(W)$ is the *prior implausibility score* of W, and $d(W, \langle i, c, \varphi\rangle)$ is the *disagreement score* for world W and $\langle i, c, \varphi\rangle$. The set of most plausible worlds $\mathcal{Y}_\sigma$ consists of those W which minimise the sum of the prior implausibility and the total disagreement with σ. Note that by summing the scores of each report $\langle i, c, \varphi\rangle$ with equal weight, we treat each report independently. This construction is informally inspired by the form of the so-called *Markovian observation systems* of Boutilier, Friedman, and Halpern (1998, Eq. (5)).

Score-based operators generalise elementary conditioning operators with **K-conjunction**.

Proposition 5.3.3. *Any elementary conditioning operator satisfying* ***K-conjunction*** *is score-based.*

Proof. Take any elementary conditioning operator corresponding to some mapping $\sigma \mapsto \langle \mathcal{X}_\sigma, \mathcal{Y}_\sigma\rangle$ and total preorder $\leq$, and suppose **K-conjunction** holds. Write

$$k(W) = |\{W' \in \mathcal{W} \mid W' \leq W\}|$$

Then we have $W \leq W'$ iff $k(W) \leq k(W')$. Set

$$r_0(W) = \begin{cases} \infty, & W \notin \mathcal{X}_\emptyset \\ k(W), & W \in \mathcal{X}_\emptyset \end{cases},$$

[13] At the very least, the case $p \in B_d^\sigma$ should not be *excluded*.

[14] In Section 5.4 we make this argument more precise by providing an impossibility result which shows conditioning operators with some basic properties cannot accept p in sequences such as this.

$$d(W, \langle i, c, \varphi \rangle) = \begin{cases} \infty, & W \notin \mathcal{X}_{\langle i,c,\varphi \rangle} \\ 0, & W \in \mathcal{X}_{\langle i,c,\varphi \rangle} \end{cases}.$$

For any sequence σ, repeated applications of **K-conjunction** (and the fact that $\mathcal{X}_\sigma$ is elementary) give $r_\sigma(W) < \infty$ iff $W \in \mathcal{X}_\sigma$. Similarly, the choice of r_0 gives $\mathrm{argmin}_{W \in \mathcal{X}_\sigma} r_\sigma(W) = \min_\leq \mathcal{X}_\sigma = \mathcal{Y}_\sigma$. Hence the operator is score-based. □

We now give a concrete example.

Definition 5.3.5. *Define a score-based operator* excess-min *by setting* $r_0(W) = 0$ *and*

$$d(W, \langle i, c, \varphi \rangle) = \begin{cases} |\Pi_i^W[\varphi] \setminus \|\varphi\||, & W, c \models \mathsf{S}_i\varphi \\ \infty, & \textit{otherwise.} \end{cases}$$

The set of possible worlds $\mathcal{X}_\sigma$ is the same as for the earlier operators. All worlds are *a priori* equiplausible according to r_0. The disagreement score d is defined as the number of propositional valuations in the "excess" of $\Pi_i^W[\varphi]$ which are not models of φ, i.e. the number of $\neg\varphi$ valuations which are indistinguishable from some φ valuation. The intuition here is that *sources tend to only report formulas on which they have expertise.* The minimum score 0 is attained exactly when i has expertise on φ; other worlds are ordered by how much they deviate from this ideal.

Again, one can verify that excess-min satisfies the basic postulates of Section 5.2.1 by an argument similar to the one employed in Proposition 5.2.4. It can also be seen that $\mathcal{X}_\sigma$ and $\mathcal{Y}_\sigma$ are elementary, and excess-min fails **Inclusion-vacuity**. It follows from Theorem 5.3.1 that excess-min is *not* a conditioning operator.[15]

Example 5.3.3. *To illustrate the differences between* excess-min *and conditioning, consider a more elaborate version of the example given at the start of this section:*

$$\sigma = (\langle i, c, p \to q \rangle, \langle j, c, p \to \neg q \rangle, \langle *, c, p \rangle, \langle i, d, p \rangle, \langle i, d, q \rangle).$$

Here the reports of i and j in case c are consistent, but inconsistent when taken with the reliable information p from $*$. Should we believe q or $\neg q$? Both our conditioning operators var-based-cond *and* part-based-cond *decline to decide, and have $[B_c^\sigma] = \mathrm{Cn}_0(p)$. However, since* excess-min *takes into account each report in the sequence, the fact that i reports both p and q in case d – and is uncontested on these reports – leads to $\mathsf{E}_i p \wedge \mathsf{E}_i q \in B_c^\sigma$. This gives $\mathsf{E}_i(p \to q) \in B_c^\sigma$ by Proposition 5.1.2 part (3), so we can make use of the report from i in case c: we have $[B_c^\sigma] = \mathrm{Cn}_0(p \wedge q)$. This example shows that score-based operators can be* more credulous *than conditioning operators (e.g. we can*

[15] We will later give an alternative proof of this fact, via an impossibility result for conditioning operators (Proposition 5.4.3).

believe $\mathsf{E}_i p$ *when* i *reports* p*), and can consequently hold stronger propositional beliefs. Indeed, the conditioning operators* var-based-cond *and* part-based-cond *are not able to make use of the reports from* i *in case* d *to form beliefs about case* c*, and in this sense do not make full use of the available information in* σ*.*

5.4 One-Step Revision

The postulates of Section 5.2.1 only set out very basic requirements for an operator. In this section we introduce some more demanding postulates which address how beliefs should change when a sequence σ is extended by a new report $\langle i, c, \varphi\rangle$. In view of **Rearrangement**, we do not view this process as *revision* of B^σ by $\langle i, c, \varphi\rangle$, but rather as *reinterpretation* of σ in light of a new report $\langle i, c, \varphi\rangle$. The postulates we introduce can therefore be seen as *coherency* requirements, which place some constraints on this reinterpretation.

First, we address how propositional beliefs should be affected by reliable information.

> **AGM-$*$.** For any σ and $c \in \mathcal{C}$ there is an AGM operator $\star$ for $[B^\sigma_c]$ such that $\left[B_c^{\sigma\cdot\langle *,c,\varphi\rangle}\right] = [B^\sigma_c] \star \varphi$ whenever $\neg\varphi \notin K^\sigma_c$

AGM-$*$ says that receiving information from the reliable source $*$ acts in accordance with the well-known AGM postulates (Alchourrón, Gärdenfors, and Makinson 1985) for propositional belief revision (provided we are not in the degenerate case where the new report φ was already *known* to be false). Since AGM revision operators are characterised by total preorders over valuations (Grove 1988; Katsuno and Mendelzon 1991), it is no surprise that our order-based constructions are consistent with **AGM-$*$**.

Proposition 5.4.1. var-based-cond*,* part-based-cond *and* excess-min *satisfy* ***AGM-$*$****.*

We require some preliminary results. For a case $c \in \mathcal{C}$ and valuation $v \in \mathcal{V}$, write $\mathcal{W}_{c\,:\,v} = \{W \in \mathcal{W} \mid v_c^W = v\}$ for the set of worlds whose c valuation is v.

Lemma 5.4.1. *For any model-based operator, sequence* σ*, case* c*, and valuation* $v \in \mathcal{V}$*,*

$$v \in \|[B^\sigma_c]\| \iff \mathcal{Y}_\sigma \cap \mathcal{W}_{c\,:\,v} \neq \emptyset$$

Proof. $\implies$: We show the contrapositive. Suppose $\mathcal{Y}_\sigma \cap \mathcal{W}_{c\,:\,v} = \emptyset$. Let ψ be any propositional formula such that $\|\psi\| = \mathcal{V} \setminus \{v\}$. Now for any $W \in \mathcal{Y}_\sigma$, we have $W \notin \mathcal{W}_{c\,:\,v}$, i.e. $v_c^W \neq v$. Hence $v_c^W \in \|\psi\|$, so $W, c \models \psi$. By definition of the belief set of a model-based operator, we have $\psi \in B^\sigma_c$. But ψ is a propositional formula, so $\psi \in [B^\sigma_c]$. Since $v \notin \|\psi\|$, we have $v \notin \|[B^\sigma_c]\|$.

$\Longleftarrow$: Suppose there is some $W \in \mathcal{Y}_\sigma \cap \mathcal{W}_{c\,:\,v}$. Let $\varphi \in [B_c^\sigma]$. Then, in particular, $\varphi \in B_c^\sigma$, so $W, c \models \varphi$ by $W \in \mathcal{Y}_\sigma$ and the definition of the model-based belief set. That is, $v = v_c^W \in \|\varphi\|$. Since $\varphi \in [B_c^\sigma]$ was arbitrary, we have $v \in \|[B_c^\sigma]\|$. ☐

We have a sufficient condition for **AGM-$*$** for score-based operators.

Lemma 5.4.2. *Suppose a score-based operator is such that for each $c \in \mathcal{C}$ and $\varphi \in \mathcal{L}_0$ there is a constant $M \in \mathbb{N}$ with*

$$d(W, \langle *, c, \varphi\rangle) = \begin{cases} M, & W, c \models \varphi \\ \infty, & W, c \models \neg\varphi \end{cases}$$

*for all W. Then **AGM-$*$** holds.*

Proof. Take a score-based operator with the stated property. Let σ be a sequence and take $c \in \mathcal{C}$. Without loss of generality, there is some $\varphi \in \mathcal{L}_0$ such that $\neg\varphi \notin K_c^\sigma$ (otherwise **AGM-$*$** trivially holds). Since any score-based operator is model-based and therefore satisfies **Closure**, we have that K^σ is inconsistent iff $K_c^\sigma = \mathcal{L}$. But since K_c^σ does not contain $\neg\varphi$, it must be the case that K^σ is consistent.

Now, set

$$k(v) = \min\{r_\sigma(W) \mid W \in \mathcal{X}_\sigma \cap \mathcal{W}_{c\,:\,v}\}$$

where $\min\emptyset = \infty$. Note that $k(v) = \infty$ if and only if $\mathcal{X}_\sigma \cap \mathcal{W}_{c\,:\,v} = \emptyset$. Then k defines a total preorder $\preceq$ on valuations, where $v \preceq v'$ iff $k(v) \leq k(v')$. Define a propositional revision operator $\star$ for $[B_c^\sigma]$ by

$$[B_c^\sigma] \star \varphi = \{\psi \in \mathcal{L}_0 \mid \min_\preceq \|\varphi\| \subseteq \|\psi\|\}$$

To show that $\star$ satisfies the AGM postulates (for $[B_c^\sigma]$) it is sufficient to show that the models of $[B_c^\sigma]$ are exactly the $\preceq$-minimal valuations.

Claim 5.4.1. $\|[B_c^\sigma]\| = \min_\preceq \mathcal{V}$.

Proof of claim. "$\subseteq$": let $v \in \|[B_c^\sigma]\|$. By Lemma 5.4.1, there is some $W \in \mathcal{Y}_\sigma \cap \mathcal{W}_{c\,:\,v}$. Since $W \in \mathcal{X}_\sigma$ too, by definition of k we have $k(v) \leq r_\sigma(W) < \infty$. Now let $v' \in \mathcal{V}$. Without loss of generality assume $k(v') < \infty$. Then there is some $W' \in \mathcal{X}_\sigma \cap \mathcal{W}_{c\,:\,v'}$ such that $k(v') = r_\sigma(W')$. But $W' \in \mathcal{X}_\sigma$ and $W \in \mathcal{Y}_\sigma$ gives $r_\sigma(W) \leq r_\sigma(W')$, so

$$k(v) \leq r_\sigma(W) \leq r_\sigma(W') = k(v')$$

i.e. $v \preceq v'$. Hence v is $\preceq$-minimal.

"$\supseteq$": let $v \in \min_\preceq \mathcal{V}$. Since K^σ is consistent, there is some $\hat{W} \in \mathcal{X}_\sigma$. Writing $\hat{v} = v_c^{\hat{W}}$, we have $\hat{W} \in \mathcal{X}_\sigma \cap \mathcal{W}_{c\,:\,\hat{v}}$, so $v \preceq \hat{v}$ implies

$$k(v) \leq k(\hat{v}) \leq r_\sigma(\hat{W}) < \infty$$

Hence there must be some $W \in \mathcal{X}_\sigma \cap \mathcal{W}_{c\,:\,v}$ such that $k(v) = r_\sigma(W)$. We claim that, in fact, $W \in \mathcal{Y}_\sigma$. Indeed, for any $W' \in \mathcal{X}_\sigma$ we have $v \preceq v_c^{W'}$, so

$$r_\sigma(W) = k(v) \leq k(v_c^{W'}) \leq r_\sigma(W')$$

That is, $W \in \mathcal{Y}_\sigma \cap \mathcal{W}_{c\,:\,v}$. By Lemma 5.4.1, $v \in \|[B_c^\sigma]\|$. ♢

So, $\star$ is indeed an AGM operator for $[B_c^\sigma]$. Now take $\varphi \in \mathcal{L}_0$ such that $\neg\varphi \notin K_c^\sigma$. Write $\rho = \sigma \cdot \langle *, c, \varphi \rangle$. We claim the following.

Claim 5.4.2. $\|[B_c^\rho]\| = \min_\preceq \|\varphi\|$.

Proof of claim. "$\subseteq$": let $v \in \|[B_c^\rho]\|$. By Lemma 5.4.1 again, there is some $W \in \mathcal{Y}_\rho \cap \mathcal{W}_{c\,:\,v}$. Since $\langle *, c, \varphi \rangle \in \rho$ and $d(W, \langle *, c, \varphi \rangle) \leq r_\rho(W) < \infty$, we must have $W, c \models \varphi$ by the assumed property of the score function d. Hence $v = v_c^W \in \|\varphi\|$.

Now since $\mathcal{Y}_\rho \subseteq \mathcal{X}_\rho$, we have $W \in \mathcal{Y}_\rho \subseteq \mathcal{X}_\rho \subseteq \mathcal{X}_\sigma$, so $W \in \mathcal{X}_\sigma \cap \mathcal{W}_{c\,:\,v}$. By definition of k, we have $k(v) \leq r_\sigma(W)$. Take any $v' \in \|\varphi\|$. Without loss of generality, assume $k(v') < \infty$, so that there is some $W' \in \mathcal{X}_\sigma \cap \mathcal{W}_{c\,:\,v'}$ with $k(v') = r_\sigma(W')$. Since $v_c^{W'} = v' \in \|\varphi\|$, we have $W', c \models \varphi$. Consequently, by the property of d again, $d(W', \langle *, c, \varphi \rangle) = M$. Since $W' \in \mathcal{X}_\sigma$ gives $r_\sigma(W') < \infty$, it follows that

$$r_\rho(W') = r_\sigma(W') + M < \infty$$

so $W' \in \mathcal{X}_\rho$. Recall that $W, c \models \varphi$ too, so $d(W, \langle *, c, \varphi \rangle) = M$ also. From $W \in \mathcal{Y}_\rho$ and $W' \in \mathcal{X}_\rho$ we get

$$\begin{aligned} r_\sigma(W) &= r_\rho(W) - M \\ &\leq r_\rho(W') - M \\ &= r_\rho(W') - d(W', \langle *, c, \varphi \rangle) \\ &= r_\sigma(W') \end{aligned}$$

This yields

$$k(v) \leq r_\sigma(W) \leq r_\sigma(W') = k(v')$$

and $v \preceq v'$ as required.

"$\supseteq$": let $v \in \min_\preceq \|\varphi\|$. Since $\neg\varphi \notin K_c^\sigma$, there is some $\hat{W} \in \mathcal{X}_\sigma$ such that $\hat{W}, c \models \varphi$. Writing $\hat{v} = v_c^{\hat{W}}$, we have $\hat{v} \in \|\varphi\|$. Hence $v \preceq \hat{v}$. This implies

$$k(v) \leq k(\hat{v}) \leq r_\sigma(\hat{W}) < \infty$$

so there must be some $W \in \mathcal{X}_\sigma \cap \mathcal{W}_{c\,:\,v}$ with $k(v) = r_\sigma(W)$. Since $v_c^W = v \in \|\varphi\|$, we have $W, c \models \varphi$. By the assumed property of d, we get $d(W, \langle *, c, \varphi \rangle) = M$. Hence

$$r_\rho(W) = r_\sigma(W) + d(W, \langle *, c, \varphi \rangle) = r_\sigma(W) + M < \infty$$

so $W \in \mathcal{X}_\rho$ too. We will show that $W \in \mathcal{Y}_\rho$. Let $W' \in \mathcal{X}_\rho$. Then we must have $d(W', \langle *, c, \varphi\rangle) = M$ and $W', c \models \varphi$. That is, $v_c^{W'} \in \|\varphi\|$. By minimality of v, we have $v \preceq v_c^{W'}$. Noting that $W' \in \mathcal{X}_\rho \subseteq \mathcal{X}_\sigma$, we get

$$r_\sigma(W) = k(v) \leq k(v_c^{W'}) \leq r_\sigma(W')$$

Consequently,

$$r_\rho(W) = r_\sigma(W) + M \leq r_\sigma(W') + M = r_\rho(W')$$

This shows $W \in \mathcal{Y}_\rho$, i.e. $\mathcal{Y}_\rho \cap \mathcal{W}_{c\,:\,v} \neq \emptyset$. By Lemma 5.4.1, we are done. $\diamondsuit$

Noting that $\|[B_c^\sigma]\| \star \varphi = \min_\preceq \|\varphi\|$, it follows from Claim 5.4.2 that $\mathrm{Cn}_0([B_c^\rho]) = \mathrm{Cn}_0([B_c^\sigma] \star \varphi)$. But $[B_c^\rho]$ is deductively closed by **Closure**, and $[B_c^\sigma] \star \varphi$ is deductively closed by construction. Hence $[B_c^\rho] = [B_c^\sigma] \star \varphi$, as required for **AGM-$*$**. $\square$

As a consequence of Proposition 5.3.3 (and the construction of d in its proof), one can apply Lemma 5.4.2 with $M = 0$ for conditioning operators with **K-conjunction** and a certain natural property.

Corollary 5.4.1. *Suppose an elementary conditioning operator satisfying* ***K-conjunction*** *has the property that*

$$W \in \mathcal{X}_{\langle *, c, \varphi\rangle} \iff W, c \models \varphi$$

Then ***AGM-$*$*** *holds.*

We can now prove Proposition 5.4.1.

Proof of Proposition 5.4.1. For the conditioning operators var-based-cond and part-based-cond, it is easily verified that the condition in Corollary 5.4.1 holds, and thus **AGM-$*$** does also. For the score-based operator excess-min, we may use Lemma 5.4.2 with $M = 0$. $\square$

Thus, we do indeed extend AGM revision in the case of reliable information. What about non-reliable information? First note that the analogue of **AGM-$*$** for ordinary sources $i \neq *$ is *not* desirable. In particular, we should not have the **Success** postulate:

$$\varphi \in B_c^{\sigma \cdot \langle i, c, \varphi\rangle}.$$

Indeed, the sequence in Example 5.2.2 with $\varphi = \neg p \wedge q$ already shows that **Success** would conflict with the basic postulates. However, there are weaker modifications of **Success** which may be more appropriate. We consider two such postulates.

Cond-success. If $\mathsf{E}_i\varphi \in B_c^\sigma$ and $\neg\varphi \notin B_c^\sigma$, then $\varphi \in B_c^{\sigma \cdot \langle i, c, \varphi\rangle}$

Strong-cond-success. If $\neg(\mathsf{E}_i\varphi \wedge \varphi) \notin B_c^\sigma$, then $\varphi \in B_c^{\sigma \cdot \langle i, c, \varphi\rangle}$

Cond-success says that if i is deemed an expert on φ, which is consistent with current beliefs, then φ is accepted after i reports it. That is, the acceptance of φ is *conditional* on prior beliefs about the expertise of i (on φ). **Strong-cond-success** weakens the antecedent by only requiring that $\mathsf{E}_i\varphi$ and φ are jointly consistent with current beliefs (i.e. i need not be considered an expert on φ). In other words, we should believe reports if there is no reason not to. It is easily shown that **Closure** and **Strong-cond-success** implies **Cond-success**. We once again revisit our examples.

Proposition 5.4.2. var-based-cond*,* part-based-cond *and* excess-min *satisfy* ***Cond-success****, and* excess-min *additionally satisfies* ***Strong-cond-success****.*

As a first step in the proof, we present sufficient conditions for conditioning operators to satisfy **Cond-success**. In fact, we do not need to impose any condition on the total preorder $\leq$: a natural constraint on the mapping $\sigma \mapsto \mathcal{X}_\sigma$ (together with some basic postulates) is enough.

Lemma 5.4.3. *Suppose an elementary conditioning operator satisfies* ***K-conjunction****,* ***Soundness*** *and*

$$W, c \models \varphi \implies W \in \mathcal{X}_{\langle i,c,\varphi\rangle}$$

Then ***Cond-success*** *holds.*

Proof. Suppose an elementary conditioning operator corresponding to the mapping $\sigma \mapsto \langle \mathcal{X}_\sigma, \mathcal{Y}_\sigma \rangle$ and total preorder $\leq$ satisfies **K-conjunction**, **Soundness** and has the stated property.

Let σ be a sequence and $c \in \mathcal{C}$. Suppose $\mathsf{E}_i\varphi \in B_c^\sigma$ and $\neg\varphi \notin B_c^\sigma$. Write $\rho = \sigma \cdot \langle i, c, \varphi\rangle$. We need to show $\varphi \in B_c^\rho$.

By $\neg\varphi \notin B_c^\sigma$, there is some $W \in \mathcal{Y}_\sigma$ such that $W, c \models \varphi$. Hence $W \in \mathcal{X}_{\langle i,c,\varphi\rangle}$. By elementariness and **K-conjunction**, we have $\mathcal{X}_\rho = \mathcal{X}_\sigma \cap \mathcal{X}_{\langle i,c,\varphi\rangle}$. Since $W \in \mathcal{Y}_\sigma \subseteq \mathcal{X}_\sigma$, we get $W \in \mathcal{X}_\rho$.

Now take any $W' \in \mathcal{Y}_\rho$. Then W' is $\leq$-minimal in $\mathcal{X}_\rho$, so $W' \leq W$. But W is $\leq$-minimal in $\mathcal{X}_\sigma$, so $W' \in \mathcal{Y}_\rho \subseteq \mathcal{X}_\rho \subseteq \mathcal{X}_\sigma$ gives $W' \in \mathcal{Y}_\sigma$ also. Consequently, $\mathsf{E}_i\varphi \in B_c^\sigma$ means $W', c \models \mathsf{E}_i\varphi$. On the other hand, **Soundness** together with $\langle i, c, \varphi\rangle \in \rho$ and $W' \in \mathcal{X}_\rho$ means $W', c \models \mathsf{S}_i\varphi$. Hence $W', c \models \mathsf{E}_i\varphi \wedge \mathsf{S}_i\varphi$. From Proposition 5.1.2 part (4), we get $W', c \models \varphi$.

We have shown that φ holds in case c at an arbitrary world in $\mathcal{Y}_\rho$. Hence $\varphi \in B_c^\rho$, as required. □

Similarly, we have a sufficient condition for score-based operators to satisfy **Strong-cond-success**: the postulate follows if worlds in which i makes an expert, truthful report are strictly more plausible than worlds in which i makes a false report.

Lemma 5.4.4. *Suppose a score-based operator is such that for any* $i \in \mathcal{S}$*,* $c \in \mathcal{C}$*,* $\varphi \in \mathcal{L}_0$ *and* $W, W' \in \mathcal{W}$*,*

$$\begin{aligned} &W, c \models \mathsf{E}_i\varphi \wedge \varphi \text{ and } W', c \models \neg\varphi \\ &\qquad \implies d(W, \langle i, c, \varphi\rangle) < d(W', \langle i, c, \varphi\rangle) \end{aligned}$$

Then ***Strong-cond-success*** *holds.*

Proof. Suppose a score-based operator has the stated property. Take σ such that $\neg(\mathsf{E}_i\varphi \wedge \varphi) \notin B^\sigma_c$. Write $\rho = \sigma \cdot \langle i, c, \varphi\rangle$. We need to show that $\varphi \in B^\rho_c$.

First note that by $\neg(\mathsf{E}_i\varphi \wedge \varphi) \notin B^\sigma_c$ and the definition of B^σ for score-based operators, there is $W \in \mathcal{Y}_\sigma$ such that $W, c \models \mathsf{E}_i\varphi \wedge \varphi$.

Take any $W' \in \mathcal{Y}_\rho$. Suppose, for the sake of contradiction, that $W', c \not\models \varphi$. Then by the hypothesised property of the score function d, we have

$$d(W, \langle i, c, \varphi\rangle) < d(W', \langle i, c, \varphi\rangle)$$

Now, $W \in \mathcal{Y}_\sigma$ and $W' \in \mathcal{Y}_\rho \subseteq \mathcal{X}_\rho \subseteq \mathcal{X}_\sigma$ gives $r_\sigma(W) \le r_\sigma(W')$. Thus

$$\begin{aligned} r_\rho(W) &= r_\sigma(W) + d(W, \langle i, c, \varphi\rangle) \\ &\le r_\sigma(W') + d(W, \langle i, c, \varphi\rangle) \\ &< r_\sigma(W') + d(W', \langle i, c, \varphi\rangle) \\ &= r_\rho(W') < \infty \end{aligned}$$

i.e. $r_\rho(W) < r_\rho(W') < \infty$. But this means $W \in \mathcal{X}_\rho$ and W' is not minimal in $\mathcal{X}_\rho$ under r_ρ, contradicting $W' \in \mathcal{Y}_\rho$. Hence $W', c \models \varphi$.

Since W' was an arbitrary member of $\mathcal{Y}_\rho$, we have shown $\varphi \in B^\rho_c$, and thus **Strong-cond-success** is shown. ☐

The main result now follows.

Proof of Proposition 5.4.2. For the conditioning operators var-based-cond and part-based-cond, **Cond-success** follows from Lemma 5.4.3 since $W, c \models \varphi$ implies $W, c \models \mathsf{S}_i\varphi$. For the score-based operator excess-min, one can easily check that the condition in Lemma 5.4.4 holds, and thus **Strong-cond-success** and **Cond-success** follow. ☐

By omission, the reader may suppose that the conditioning operators fail **Strong-cond-success**. This is correct, and we can in fact say even more: *no* conditioning operator with a few basic properties – all of which are satisfied by var-based-cond and part-based-cond – can satisfy **Strong-cond-success**. In what follows, for a permutation $\pi : \mathcal{S} \to \mathcal{S}$ with $\pi(*) = *$, write $\pi(W)$ for the world with $v_c^{\pi(W)} = v_c^W$ and $\Pi_i^{\pi(W)} = \Pi^W_{\pi(i)}$. We have an impossibility result.

Proposition 5.4.3. *No elementary conditioning operator satisfying the basic postulates can simultaneously satisfy the following properties:*

1. $K^\emptyset = \mathrm{Cn}(\emptyset)$

2. *If π is a permutation of $\mathcal{S}$ with $\pi(*) = *$, $W \simeq \pi(W)$*

3. ***Refinement***

4. ***Strong-cond-success***

However, any proper subset of (1) - (4) is satisfiable.

(1) says that before any reports are received, we only know tautologies. As remarked earlier, this is not an *essential* property, but is reasonable when no prior knowledge is available. (2) is an anonymity postulate: it says that permuting the "names" of sources does not affect the plausibility of a world, and is a desirable property in light of (1). **Refinement**, introduced in Section 5.3.1, says that worlds in which all sources have more expertise are preferred.

Proof. Take distinct sources $i_1, i_2 \in \mathcal{S} \setminus \{*\}$, distinct cases $c, d \in \mathcal{C}$, and distinct valuations $v_1, v_2 \in \mathcal{V}$. Let $\varphi_1, \varphi_2 \in \mathcal{L}_0$ be propositional formulas with $\|\varphi_k\| = \{v_k\}$ $(k \in \{1, 2\})$. Suppose for contradiction that some elementary conditioning operator – satisfying the basic postulates – has the stated properties.

Define a sequence

$$\sigma = (\langle *, c, \varphi_1 \vee \varphi_2 \rangle, \langle i_1, c, \varphi_1 \rangle, \langle i_2, c, \varphi_2 \rangle).$$

Let $\Pi_\perp$ denote the unit partition $\{\{u\} \mid u \in \mathcal{V}\}$, and let $\widehat{\Pi}$ denote the partition

$$\{\{v_1, v_2\}\} \cup \{\{u\} \mid u \in \mathcal{V} \setminus \{v_1, v_2\}\},$$

i.e. the partition obtained from $\Pi_\perp$ by merging the cells of v_1 and v_2. Consider worlds W_1, W_2 given by

$$\begin{aligned} v_{c'}^{W_k} &= v_k \qquad (c' \in \mathcal{C}) \\ \Pi_i^{W_k} &= \begin{cases} \widehat{\Pi}, & (k = 1 \text{ and } i = i_2) \text{ or } (k = 2 \text{ and } i = i_1) \\ \Pi_\perp, & \text{otherwise} \end{cases} \end{aligned}$$

That is, W_1 has v_1 as its valuation for all cases, i_2 has partition $\widehat{\Pi}$, and all other sources have the finest partition $\Pi_\perp$; similarly W_2 has v_2 for its valuations and all sources except i_1 have $\Pi_\perp$.

Let $\leq$ denote the total preorder associated with the conditioning operator.

Claim 5.4.3. *$W_1 \simeq W_2$.*

Proof of claim. Let π be the permutation of $\mathcal{S}$ which swaps i_1 and i_2. It is easily observed that $\pi(W_1)$ is partition-equivalent to W_2. By reflexivity of partition refinement, $\pi(W_1) \preceq W_2$ and $W_2 \preceq \pi(W_1)$. By **Refinement**, we get $\pi(W_1) \simeq W_2$. By property (2), $W_1 \simeq \pi(W_1)$. By transitivity of $\simeq$ we get $W_1 \simeq W_2$ as desired. $\diamondsuit$

Now, from the basic postulates, property (1) and Proposition 5.2.1 we have $K^\sigma = \mathrm{Cn}(G^\sigma_{\mathsf{snd}})$. By elementariness and Lemma 5.3.2, we get $\mathcal{X}_\sigma = \mathrm{mod}(K^\sigma) = \mathrm{mod}(G^\sigma_{\mathsf{snd}})$. It is easily checked that both W_1 and W_2 satisfy the soundness statements corresponding to σ, and thus $W_1, W_2 \in \mathrm{mod}(G^\sigma_{\mathsf{snd}}) = \mathcal{X}_\sigma$.

Claim 5.4.4. $W_1, W_2 \in \mathcal{Y}_\sigma$.

Proof of claim. We show W_1 and W_2 are $\leq$-minimal in $\mathcal{X}_\sigma$. Take any $W \in \mathcal{X}_\sigma$. Then $W \in \mathrm{mod}(G^\sigma_{\mathsf{snd}})$, so $W, c \models \mathsf{S}_*(\varphi_1 \vee \varphi_2)$, i.e. $v^W_c \in \{v_1, v_2\}$. We consider two cases.

- **Case 1** ($v^W_c = v_1$). By $W \in \mathrm{mod}(G^\sigma_{\mathsf{snd}})$ again we have $W, c \models \mathsf{S}_{i_2}\varphi_2$, i.e

$$v_1 = v^W_c \in \Pi^W_{i_2}[\varphi_2] = \Pi^W_{i_2}[v_2].$$

  It follows that $\{v_1, v_2\} \subseteq \Pi^W_{i_2}[v_2]$, and that $\widehat{\Pi}$ refines $\Pi^W_{i_2}$. Since $\widehat{\Pi}$ is the partition of i_2 in W_1, and all other sources have the finest partition $\Pi_\perp$, we get $W_1 \preceq W$. By **Refinement**, $W_1 \leq W$. Since $W_1 \simeq W_2$ we have $W_2 \leq W$ also.

- **Case 2** ($v^W_c = v_2$). Applying a near-identical argument to that used in case 1 with soundness of the report $\langle i_1, c, \varphi_1 \rangle$, we get $W_1, W_2 \leq W$.

In either case, both $W_1 \leq W$ and $W_2 \leq W$, so $W_1, W_2 \in \mathcal{Y}_\sigma$. $\diamondsuit$

Now we consider case d. Since

$$W_1, d \models \mathsf{E}_{i_1}\varphi_1 \wedge \varphi_1$$

and $W_1 \in \mathcal{Y}_\sigma$, $\neg(\mathsf{E}_{i_1}\varphi_1 \wedge \varphi_1) \notin B^\sigma_d$. Writing $\rho = \sigma \cdot \langle i_1, d, \varphi_1 \rangle$, we get from **Strong-cond-success** that $\varphi_1 \in B^\rho_d$.

Note that $W_2, d \models \mathsf{S}_{i_1}\varphi_1$, so $W_2 \in \mathrm{mod}(G^\rho_{\mathsf{snd}}) = \mathrm{mod}(K^\rho) = \mathcal{X}_\rho$. Since W_2 is $\leq$-minimal in $\mathcal{X}_\sigma$ and

$$X_\rho = \mathrm{mod}(G^\rho_{\mathsf{snd}}) \subseteq \mathrm{mod}(G^\sigma_{\mathsf{snd}}) = \mathcal{X}_\sigma,$$

W_2 is also $\leq$-minimal in $\mathcal{X}_\rho$, i.e. $W_2 \in \mathcal{Y}_\rho$. Now $\varphi_1 \in B^\rho_d$ gives $W_2, d \models \varphi_1$. Since $v^{W_2}_d = v_2$ and $\|\varphi_1\| = \{v_1\}$, this means $v_1 = v_2$. But v_1 and v_2 were assumed to be distinct: contradiction. $\square$

Proposition 5.4.3 highlights an important difference between conditioning and score-based operators, and hints that a fixed plausibility order may be too restrictive: we need to allow the order to be responsive to new reports in order to satisfy properties such as **Strong-cond-success**.

To further this point, two of the postulates involved in the characterisation of elementary conditioning operators – **Duplicate-removal** and **Inclusion-vacuity** – are already enough on their own to imply a somewhat questionable property when combined with **Strong-cond-success**.

Decisiveness. If $\langle i, c, \varphi\rangle \in \sigma$ then either $\varphi \in B_c^\sigma$ or $\neg \mathsf{E}_i\varphi \in B_c^\sigma$.

This property says that each report is either accepted, or the reporting source is distrusted. Thus, no room is left for an operator to abstain on a particular report. It can be easily seen that **Decisiveness** fails for var-based-cond, part-based-cond and excess-min by considering the simple sequence $\sigma = (\langle i, c, p\rangle, \langle j, c, \neg p\rangle)$, and indeed we argue this is the intuitively "correct" behaviour. The tension between **Strong-cond-success** and **Duplicate-removal** – which says that duplicate reports can be removed without effecting beliefs – is already evident from this example when one considers $\sigma \cdot \langle i, c, p\rangle$. Formally, we have the following.

Proposition 5.4.4.

1. *The basic postulates,* ***Duplicate-removal*** *and* ***Strong-cond-success*** *imply* ***Decisiveness****.*

2. *The basic postulates,* ***Inclusion-vacuity*** *and* ***Strong-cond-success*** *imply* ***Decisiveness****.*

Proof.

1. Suppose $\langle i, c, \varphi\rangle \in \sigma$. First suppose $\neg(\mathsf{E}_i\varphi \land \varphi) \notin B_c^\sigma$. Then **Strong-cond-success** gives $\varphi \in B_c^{\sigma\cdot\langle i,c,\varphi\rangle}$. But since $\langle i, c, \varphi\rangle$ already appears in σ, **Rearrangement** and **Duplicate-removal** give $B^{\sigma\cdot\langle i,c,\varphi\rangle} = B^\sigma$. Thus $\varphi \in B_c^\sigma$.

   Now suppose $\neg(\mathsf{E}_i\varphi \land \varphi) \in B_c^\sigma$. We claim $\neg\mathsf{E}_i\varphi \in B_c^\sigma$, i.e. $W, c \models \neg\mathsf{E}_i\varphi$ for all $W \in \mathcal{Y}_\sigma$. Indeed, take any $W \in \mathcal{Y}_\sigma$. Then $W, c \models \neg(\mathsf{E}_i\varphi \land \varphi)$, so either $W, c \models \neg\mathsf{E}_i\varphi$ or $W, c \models \neg\varphi$. In the former case we are done. In the latter case, an application of **Soundness** and **Containment** gives $W, c \models \mathsf{S}_i\varphi$, so in fact $W, c \models \mathsf{S}_i\varphi \land \neg\varphi$. But $\mathsf{S}_i\varphi \land \neg\varphi \to \neg\mathsf{E}_i\varphi$ is a validity of the logic (Proposition 5.1.2 (4)), so $W, c \models \neg\mathsf{E}_i\varphi$ also. Hence $\neg\mathsf{E}_i\varphi \in B_c^\sigma$, and **Decisiveness** is shown.

2. By Lemma 5.3.1 the basic postulates and **Inclusion-vacuity** already imply **Duplicate-removal**, so we may conclude by (1).

□

5.5 Selective Change

In the previous section we saw how a single formula φ may be accepted when it is received as an additional report. But what can we say about propositional beliefs when taking into account the *whole sequence* σ? To investigate this we introduce an analogue of *selective revision* (Fermé and Hansson 1999), in which propositional beliefs are formed by "selecting" only a part of each input report (e.g., some part consistent with the source's expertise). For example, in Example 5.3.1 we saw that when given $\sigma = (\langle *, c, p\rangle, \langle i, c, \neg p \wedge q\rangle)$, var-based-cond outputs propositional beliefs $[B_c^\sigma] = \mathrm{Cn}_0(p \wedge q)$. Intuitively, the report from $*$ is taken as-is, whereas the report of $\neg p \wedge q$ from i is weakened to just q. The resulting formulas are combined conjunctively to form the propositional belief set. We formalise this idea via *selection schemes*. In what follows, write $\sigma \upharpoonright c = \{\langle i, \varphi\rangle \mid \langle i, c, \varphi\rangle \in \sigma\}$ for the c-reports in σ.

Definition 5.5.1. *A* selection scheme *is a mapping f assigning to each $*$-consistent sequence σ a function $f_\sigma : \mathcal{S} \times \mathcal{C} \times \mathcal{L}_0 \to \mathcal{L}_0$ such that $f_\sigma(i, c, \varphi) \in \mathrm{Cn}_0(\varphi)$. An operator is* selective *if there is a selection scheme f such that for all $*$-consistent σ and $c \in \mathcal{C}$,*

$$[B_c^\sigma] = \mathrm{Cn}_0(\{f_\sigma(i, c, \varphi) \mid \langle i, \varphi\rangle \in \sigma \upharpoonright c\}).$$

Thus, an operator is selective if its propositional beliefs in case c are formed by weakening each c-report and taking their consequences. Note that for $\sigma = \emptyset$ we get $[B_c^\sigma] = \mathrm{Cn}_0(\emptyset)$, so selectivity already rules out non-tautological prior propositional beliefs. Also note that in the presence of **Closure**, **Containment** and **Soundness**, selectivity implies that $[B_c^\sigma] = [B_c^\rho]$, where ρ is obtained by replacing each report $\langle i, c, \varphi\rangle$ with $\langle *, c, f_\sigma(i, c, \varphi)\rangle$.

Selectivity can be characterised by a natural postulate placing an upper bound on the propositional part of B_c^σ. For any sequence σ and case c, write $\Gamma_c^\sigma = \{\varphi \in \mathcal{L}_0 \mid \exists i \in \mathcal{S} : \langle i, \varphi\rangle \in \sigma \upharpoonright c\}$.

Boundedness. If σ is $*$-consistent, $[B_c^\sigma] \subseteq \mathrm{Cn}_0(\Gamma_c^\sigma)$

Boundedness says that the propositional beliefs in case c should not go beyond the consequences of the formulas reported in case c. In some sense this can be seen as an iterated version of **Inclusion** from AGM revision, in the case where $\left[B_c^\emptyset\right] = \mathrm{Cn}_0(\emptyset)$. We have the following characterisation.

Theorem 5.5.1. *A model-based operator is selective if and only if it satisfies **Boundedness**.*

Proof. "if": Suppose a model-based operator satisfies **Boundedness**. Take any $*$-consistent σ. For $c \in \mathcal{C}$, set

$$M_c = \|[B_c^\sigma]\|.$$

By **Boundedness**, we have $M_c \supseteq \|\Gamma_c^\sigma\|$. Now set

$$F_\sigma(i, c, \varphi) = \|\varphi\| \cup M_c.$$

Define a selection function f_σ by letting $f_\sigma(i, c, \varphi)$ be any formula with $\|f_\sigma(i, c, \varphi)\| = F_\sigma(i, c, \varphi)$. Since $F_\sigma(i, c, \varphi)$ contains the models of φ, clearly $f_\sigma(i, c, \varphi) \in \mathrm{Cn}_0(\varphi)$. Therefore f is indeed a selection function.

We claim that, for any $c \in \mathcal{C}$,

$$M_c = \bigcap_{\langle i,\varphi\rangle \in \sigma \restriction c} F_\sigma(i,c,\varphi).$$

The "$\subseteq$" inclusion is clear since, by definition, $F_\sigma(i,c,\varphi) \supseteq M_c$. For the "$\supseteq$" inclusion, suppose for contradiction that there is some $v \in \bigcap_{\langle i,\varphi\rangle \in \sigma \restriction c} F_\sigma(i,c,\varphi)$ with $v \notin M_c$.

Take any $\varphi \in \Gamma_c^\sigma$. Then there is $i \in \mathcal{S}$ such that $\langle i,\varphi\rangle \in \sigma \restriction c$, and hence $v \in F_\sigma(i,c,\varphi)$. But $v \notin M_c$ by assumption, so $v \in \|\varphi\|$. This shows $v \in \|\Gamma_c^\sigma\|$. But $\|\Gamma_c^\sigma\| \subseteq M_c$ by **Boundedness**, so $v \in M_c$; contradiction.

From this we get

$$\begin{aligned}
\|[B_c^\sigma]\| &= M_c \\
&= \bigcap_{\langle i,\varphi\rangle \in \sigma \restriction c} F_\sigma(i,c,\varphi) \\
&= \bigcap_{\langle i,\varphi\rangle \in \sigma \restriction c} \|f_\sigma(i,c,\varphi\|) \\
&= \|\{f_\sigma(i,c,\varphi) \mid \langle i,\varphi\rangle \in \sigma \restriction c\}\|
\end{aligned}$$

Since $[B_c^\sigma]$ is deductively closed (by **Closure**, which holds for all model-based operators), we get

$$[B_c^\sigma] = \mathrm{Cn}_0\left(\{f_\sigma(i,c,\varphi) \mid \langle i,\varphi\rangle \in \sigma \restriction c\}\right)$$

as required for selectivity.

"only if": Suppose a model-based operator is selective according to some selection scheme f. Take any $*$-consistent σ and $c \in \mathcal{C}$. Write

$$\Delta = \{f_\sigma(i,c,\varphi) \mid \langle i,\varphi\rangle \in \sigma \restriction c\}.$$

so that $[B_c^\sigma] = \mathrm{Cn}_0(\Delta)$. For $\langle i,\varphi\rangle \in \sigma \restriction c$ we have $f_\sigma(i,c,\varphi) \in \mathrm{Cn}_0(\varphi) \subseteq \mathrm{Cn}_0(\Gamma_c^\sigma)$ from the definition of a selection scheme and the fact that $\varphi \in \Gamma_c^\sigma$. Hence $\Delta \subseteq \mathrm{Cn}_0(\Gamma_c^\sigma)$, so

$$[B_c^\sigma] = \mathrm{Cn}_0(\Delta) \subseteq \mathrm{Cn}_0(\mathrm{Cn}_0(\Gamma_c^\sigma)) = \mathrm{Cn}_0(\Gamma_c^\sigma)$$

as required for **Boundedness**. □

The characterisation in Theorem 5.5.1 allows us to easily analyse when conditioning and score-based operators are selective. In the case of conditioning operators with $K^\emptyset = \mathrm{Cn}(\emptyset)$, we in fact have a precise characterisation. First, some terminology: say that a world W *refines* W' *at* c if for all $i \in \mathcal{S}$ we have $\Pi_i^W[v_c^W] \subseteq \Pi_i^{W'}[v_c^{W'}]$. Intuitively, this means each source is more knowledgable in case c in world W than they are in W'. Recall that $\mathcal{W}_{c\,:\,v} = \{W \in \mathcal{W} \mid v_c^W = v\}$ denotes the set of worlds whose c valuation is v. We have the following.

Proposition 5.5.1. *Suppose an elementary conditioning operator satisfies the basic postulates and has $K^{\emptyset} = \mathrm{Cn}(\emptyset)$. Then it is selective if and only if for all W, c, v there is $W' \in \mathcal{W}_{c\,:\,v}$ such that $W' \leq W$ and W refines W' at all cases $d \neq c$.*

While the condition on $\leq$ in Proposition 5.5.1 is somewhat technical, it is implied by the very natural *partition-equivalence* property from Section 5.1. Consequently, var-based-cond and part-based-cond are selective. For the score-based operator excess-min, one can show **Boundedness** holds directly using a property of the disagreement scoring function d similar to the property of $\leq$ above. Consequently, excess-min is also selective.

To prove Proposition 5.5.1, we first state some preliminary results.

Lemma 5.5.1. *Suppose W refines W' at c. Then for any $i \in \mathcal{S}$ and $\varphi \in \mathcal{L}_0$,*

$$W, c \models \mathsf{S}_i\varphi \implies W', c \models \mathsf{S}_i\varphi.$$

Proof. Suppose $W, c \models \mathsf{S}_i\varphi$. Then $v_c^W \in \Pi_i^W[\varphi]$, i.e. $\|\varphi\| \cap \Pi_i^W[v_c^W] \neq \emptyset$. By refinement, $\Pi_i^W[v_c^W] \subseteq \Pi_i^{W'}[v_c^{W'}]$. Hence $\|\varphi\| \cap \Pi_i^{W'}[v_c^{W'}] \neq \emptyset$, so $v_c^{W'} \in \Pi_i^{W'}[\varphi]$. That is, $W', c \models \mathsf{S}_i\varphi$. $\square$

Lemma 5.5.2. *For any $W \in \mathcal{W}$ and $c \in \mathcal{C}$, there is a $*$-consistent sequence σ – containing only reports for case c – such that for all $W' \in \mathcal{W}$,*

$$W' \in \mathrm{mod}(G^{\sigma}_{\mathsf{snd}}) \iff W \text{ refines } W' \text{ at } c.$$

Proof. For a valuation $v \in \mathcal{V}$, let $\varphi(v)$ be a propositional formula such that $\|\varphi(v)\| = \{v\}$. Take σ to be any enumeration of reports of the form

$$\langle i, c, \varphi(v) \rangle,$$

where $i \in \mathcal{S}$ and $v \in \Pi_i^W[v_c^W]$. Note that such a sequence exists since there are only finitely many sources and valuations. Clearly σ contains only c-reports. Since Π_*^W is the unit partition, the only report from $*$ is $\langle *, c, \varphi(v_c^W) \rangle$. Hence σ is $*$-consistent. We show the desired equivalence.

$\implies$: Suppose $W' \in \mathrm{mod}(G^{\sigma}_{\mathsf{snd}})$. Take any $i \in \mathcal{S}$. We need to show $\Pi_i^W[v_c^W] \subseteq \Pi_i^{W'}[v_c^{W'}]$. Take $v \in \Pi_i^W[v_c^W]$. By construction of σ, $\langle i, c, \varphi(v) \rangle \in \sigma$. Hence $W', c \models \mathsf{S}_i\varphi(v)$, i.e. $v_c^{W'} \in \Pi_i^{W'}[\varphi(v)] = \Pi_i^{W'}[v]$. This shows $v \in \Pi_i^{W'}[v_c^{W'}]$ as required.

$\impliedby$: Suppose W refines W' at c. Take any $\langle i, c, \varphi(v) \rangle \in \sigma$. Then $v \in \Pi_i^W[v_c^W]$, so $v_c^W \in \Pi_i^W[v] = \Pi_i^W[\varphi(v)]$. This shows $W, c \models \mathsf{S}_i\varphi(v)$, and Lemma 5.5.1 gives $W', c \models \mathsf{S}_i\varphi(v)$. Hence $W' \in \mathrm{mod}(G^{\sigma}_{\mathsf{snd}})$. $\square$

Proof of Proposition 5.5.1. Take an elementary conditioning operator with the basic postulates and $K^{\emptyset} = \mathrm{Cn}(\emptyset)$.

"if": Suppose the stated property holds. Since all conditioning operators are model-based, by Theorem 5.5.1 it suffices to show **Boundedness**. To that end, let σ be $*$-consistent and take $c \in \mathcal{C}$. We need $[B^\sigma_c] \subseteq \mathrm{Cn}_0(\Gamma^\sigma_c)$; or equivalently, by **Closure**, $\|[B^\sigma_c]\| \supseteq \|\Gamma^\sigma_c\|$.

Take any $v \in \|\Gamma^\sigma_c\|$. Since σ is $*$-consistent, B^σ is consistent by **Consistency**. Hence $\mathcal{Y}_\sigma \neq \emptyset$. Take any $W \in \mathcal{Y}_\sigma$. By the property in the statement of the result, there is $W' \in \mathcal{W}_{c\,:\,v}$ such that $W' \leq W$ and W refines W' at all cases $d \neq c$.

We claim $W' \in \mathcal{X}_\sigma$. By Proposition 5.2.1, elementariness and Lemma 5.3.2, we have $\mathcal{X}_\sigma = \mathrm{mod}(K^\sigma) = \mathrm{mod}(G^\sigma_{\mathsf{snd}})$. Take any $\langle i, d, \varphi\rangle \in \sigma$. We consider cases.

- **Case 1** ($d = c$). Here $\langle i, \varphi\rangle \in \sigma \restriction c$, so $\varphi \in \Gamma^\sigma_c$. Hence $v \in \|\Gamma^\sigma_c\| \subseteq \|\varphi\|$. Since $W' \in \mathcal{W}_{c\,:\,v}$, v is the c-valuation of W'. Hence $W', c \models \varphi$, and $W', c \models \mathsf{S}_i\varphi$ follows.

- **Case 2** ($d \neq c$). By assumption, W refines W' at d. Since $W \in \mathcal{Y}_\sigma \subseteq \mathcal{X}_\sigma$, we have $W, d \models \mathsf{S}_i\varphi$. By Lemma 5.5.1, $W', d \models \mathsf{S}_i\varphi$ also.

We have shown $W' \in \mathrm{mod}(G^\sigma_{\mathsf{snd}}) = \mathcal{X}_\sigma$. Now recall that $W \in \mathcal{Y}_\sigma$ – so W is $\leq$-minimal in $\mathcal{X}_\sigma$ – and $W' \leq W$. Thus W' is also $\leq$-minimal in $\mathcal{X}_\sigma$, i.e. $W' \in \mathcal{Y}_\sigma$. Since $W' \in \mathcal{W}_{c\,:\,v}$ also, we have by Lemma 5.4.1 that $v \in \|[B^\sigma_c]\|$, as required.

"only if": Suppose our operator is selective, i.e. satisfies **Boundedness**. To show the desired property holds, take any W, c and v. Enumerate $\mathcal{C} \setminus \{c\}$ as $\{d_1, \ldots, d_N\}$. By Lemma 5.5.2, for each $1 \leq n \leq N$ there is a $*$-consistent sequence σ_n such that

$$\mathrm{mod}(G^{\sigma_n}_{\mathsf{snd}}) = \{W' \in \mathcal{W} \mid W \text{ refines } W' \text{ at } d_n\}.$$

Now, let φ and ψ be formulas with $\|\varphi\| = \{v\}$ and $\|\psi\| = \{v^W_c\}$. Let ρ be the concatenation

$$\rho = \sigma_1 \cdots \sigma_n \cdot \langle *, c, \varphi \vee \psi\rangle.$$

Note that ρ is $*$-consistent, since each σ_n is (and only refers to case d_n). We may therefore apply **Boundedness** for case c. Taking models of both sides yields

$$\|[B^\rho_c]\| \supseteq \|\Gamma^\rho_c\| = \|\varphi \vee \psi\| = \{v, v^W_c\}.$$

In particular, $v \in \|[B^\rho_c]\|$. By Lemma 5.4.1, there is some $W' \in \mathcal{Y}_\rho \cap \mathcal{W}_{c\,:\,v}$.

We show W' has the required properties. First note that since W refines itself at each d_n, we have $W \in \mathrm{mod}(G^{\sigma_n}_{\mathsf{snd}})$. Clearly $W, c \models \psi$, so $W, c \models \mathsf{S}_*(\varphi \vee \psi)$ too. Thus $W \in \mathrm{mod}(G^\rho_{\mathsf{snd}}) = \mathcal{X}_\rho$ (using $K^\emptyset = \mathrm{Cn}(\emptyset)$). Since $W' \in \mathcal{Y}_\rho = \min_\leq \mathcal{X}_\rho$, we get $W' \leq W$ as required.

Next, take any case $d \neq c$. Then there is some n such that $d = d_n$. Since $W' \in \mathcal{Y}_\rho \subseteq \mathcal{X}_\rho = \mathrm{mod}(G^\rho_{\mathsf{snd}}) \subseteq \mathrm{mod}(G^{\sigma_n}_{\mathsf{snd}})$, we get that W refines W' at d. This completes the proof. □

5.5.1 Case Independence

In the definition of a selection scheme, we allow $f_\sigma(i, c, \varphi)$ to depend on the case c. If one views $f_\sigma(i, c, \varphi)$ as a weakening of φ which accounts for the lack of expertise of i, this is somewhat at odds with other aspects of the framework, where expertise is independent of case. For this reason it is natural to consider *case independent* selective schemes.

Definition 5.5.2. *A selection scheme f is* case independent *if $f_\sigma(i, c, \varphi) \equiv f_\sigma(i, d, \varphi)$ for all $*$-consistent σ and $i \in \mathcal{S}$, $c, d \in \mathcal{C}$ and $\varphi \in \mathcal{L}_0$.*

Say an operator is *case-independent-selective* if it is selective according to some case independent scheme. This stronger notion of selectivity can again be characterised by a postulate which bounds propositional beliefs. For any set of cases $H \subseteq \mathcal{C}$, sequence σ and $c \in \mathcal{C}$, write

$$\Gamma_c^{\sigma,H} = \{\varphi \in \mathcal{L}_0 \mid \exists i \in \mathcal{S} : \langle i, \varphi\rangle \in \sigma \restriction c \text{ and } \forall d \in H : \langle i, \varphi\rangle \notin \sigma \restriction d\}.$$

H-Boundedness. For any $*$-consistent σ, $H \subseteq \mathcal{C}$ and $c \in \mathcal{C}$,

$$[B_c^\sigma] \subseteq \mathrm{Cn}_0\left(\Gamma_c^{\sigma,H} \cup \bigcup_{d \in H} [B_d^\sigma]\right)$$

Note that **Boundedness** is obtained as the special case where $H = \emptyset$. We illustrate with an example.

Example 5.5.1. *Consider case c in the following sequence:*

$$\sigma = (\langle i, c, p\rangle, \langle j, c, q\rangle, \langle j, d, q\rangle, \langle k, d, r\rangle)$$

***Boundedness** requires that $[B_c^\sigma] \subseteq \mathrm{Cn}_0(\{p, q\})$. However, the instance of **H-Boundedness** with $H = \{d\}$ makes use of the fact that j reports q in both cases c and d, and requires $[B_c^\sigma] \subseteq \mathrm{Cn}_0(\{p\} \cup [B_d^\sigma])$. This also has an interesting implication for case d: if $\varphi \in [B_c^\sigma]$, then $p \to \varphi \in [B_d^\sigma]$. This follows since $\beta \in \mathrm{Cn}_0(\{\alpha\} \cup \Gamma)$ iff $\alpha \to \beta \in \mathrm{Cn}_0(\Gamma)$ for $\alpha, \beta \in \mathcal{L}_0$. Intuitively, this says that if p (from i) and q (from j) is enough to accept φ in case c, then φ is accepted in case d* if p is*, given that the report of q from j is repeated for d.*

The characterisation is as follows.

Theorem 5.5.2. *A model-based operator is case-independent-selective if and only if it satisfies **H-Boundedness**.*

Proof. "only if": Suppose a model-based operator is selective according to some case-independent scheme f. Take any $*$-consistent σ, $H \subseteq \mathcal{C}$ and $c \in \mathcal{C}$. For any case d, write $M_d = \|[B_d^\sigma]\|$. Note that with c_0 an arbitrary fixed case, and writing $F_\sigma(i,\varphi) = \|f_\sigma(i,c_0,\varphi\|)$, we have by case-independent-selectivity that

$$M_d = \bigcap_{\langle i,\varphi\rangle\in\sigma\restriction d} F_\sigma(i,\varphi).$$

By closure, it is sufficient for **H-Boundedness** to show that

$$M_c \supseteq \|\Gamma_c^{\sigma,H}\| \cap \bigcap_{d\in H} M_d. \tag{5.3}$$

Take any v in the set on the right-hand side. To show $v \in M_c$, take any $\langle i,\varphi\rangle \in \sigma \restriction c$. If $\varphi \in \Gamma_c^{\sigma,H}$, then clearly

$$\begin{aligned} v &\in \|\Gamma_c^{\sigma,H}\| \\ &\subseteq \|\varphi\| \\ &\subseteq \|f_\sigma(i,c,\varphi\|) \\ &= F_\sigma(i,\varphi) \end{aligned}$$

(where we use $f_\sigma(i,c,\varphi) \in \mathrm{Cn}_0(\varphi)$). Otherwise, $\varphi \notin \Gamma_c^{\sigma,H}$. Since $\langle i,\varphi\rangle \in \sigma \restriction c$, this means there is $d \in H$ such that $\langle i,\varphi\rangle \in \sigma \restriction d$. Hence $v \in M_d$ gives $v \in F_\sigma(i,\varphi)$. This shows the inclusion in (5.3), and we are done.

"if": Suppose a model-based operator satisfies **H-Boundedness**. Let σ be a $*$-consistent sequence. As before, write M_c for $\|[B_c^\sigma]\|$. For $i \in \mathcal{S}$ and $c \in \mathcal{C}$, write

$$\mathcal{C}(i,\varphi) = \{c \in \mathcal{C} \mid \langle i,\varphi\rangle \in \sigma \restriction c\},$$

and set

$$F_\sigma(i,\varphi) = \|\varphi\| \cup \bigcup_{c\in\mathcal{C}(i,\varphi)} M_c.$$

Define f by letting $f_\sigma(i,c,\varphi)$ be any propositional formula with $\|f_\sigma(i,c,\varphi\|) = F_\sigma(i,\varphi)$. Then f is a case-independent selection scheme. We show our operator is selective according to f; by closure of $[B_c^\sigma]$ for each c, it suffices to show

$$M_c = \bigcap_{\langle i,\varphi\rangle\in\sigma\restriction c} F_\sigma(i,\varphi).$$

Fix c. For the left-to-right inclusion, suppose $v \in M_c$. Take any $\langle i,\varphi\rangle \in \sigma \restriction c$. Then $c \in \mathcal{C}(i,\varphi)$, so $F_\sigma(i,\varphi) \supseteq M_c$ and thus $v \in F_\sigma(i,\varphi)$ as required.

For the right-to-left inclusion, suppose v lies in the intersection. Set

$$H = \{d \in \mathcal{C} \mid v \in M_d\}.$$

Applying **H-Boundedness** and taking the models of both sides, we obtain

$$M_c \supseteq \|\Gamma_c^{\sigma,H}\| \cap \bigcap_{d\in H} M_d. \tag{5.4}$$

Clearly $v \in \bigcap_{d \in H} M_d$ by definition of H. Let $\varphi \in \Gamma_c^{\sigma,H}$. Then there is $i \in \mathcal{S}$ such that $\langle i, \varphi \rangle \in \sigma \upharpoonright c$, and consequently $v \in F_\sigma(i, \varphi)$. We claim $v \in \|\varphi\|$. If not, by definition of $F_\sigma(i, \varphi)$ we must have $v \in \bigcup_{d \in \mathcal{C}(i,\varphi)} M_d$, i.e. there is $d \in \mathcal{C}$ such that $\langle i, \varphi \rangle \in \sigma \upharpoonright d$ and $v \in M_d$. On the one hand, $\varphi \in \Gamma_c^{\sigma,H}$ implies $d \notin H$. On the other, $v \in M_d$ gives $d \in H$ directly by the definition of H: contradiction. This shows $v \in \|\varphi\|$. Since φ was arbitrary, we have $v \in \|\Gamma_c^{\sigma,H}\|$. By (5.4) we get $v \in M_c$, and the proof is complete. □

The question of whether our concrete operators satisfy **H-Boundedness** (equivalently, whether they are case-independent-selective) is still open.

5.5.2 Expertise and Selectivity

In the existing literature on selective belief change (e.g. (Fermé and Hansson 1999; Booth and Hunter 2018)), the selection function typically acts as a means to separate out the part of new information on which the reporting sources is *credible*, or *trusted*. For instance, Booth and Hunter (2018) use a partition Π to represent an agent's perception of the incoming source's expertise; a report of φ is then weakened to $\Pi[\varphi]$ – on which the source is trusted to have expertise – before revision takes place. In our framework, an analogous connection between the selection function and trust can be captured as follows.

Definition 5.5.3. *A selection scheme f is* expertise-compatible *(EC) with an operator $\sigma \mapsto \langle B^\sigma, K^\sigma \rangle$ if for all $*$-consistent σ and $\langle i, c, \varphi \rangle \in \sigma$,*

$$\mathsf{E}_i f_\sigma(i, c, \varphi) \in B_c^\sigma.$$

That is, i is trusted on the weakened report $f_\sigma(i, c, \varphi)$ whenever i reports φ in case c in σ. Say an operator is *EC-selective* if it is selective according to some expertise-compatible scheme. While EC-selectivity may appear natural on first glance, we argue that it can be overly restrictive when expertise is derived from the input sequence itself. For example, consider the sequence

$$\sigma = (\langle i, c, p \rangle, \langle j, c, p \rangle, \langle i, d, p \rangle, \langle j, d, \neg p \rangle)$$

By **Soundness** and **Closure**, we cannot have both $\mathsf{E}_i p$ and $\mathsf{E}_j p$ in B_c^σ. Ideas of symmetry suggest that neither can we pick one of i or j over the other, so that in fact it is reasonable to have neither $\mathsf{E}_i p$ nor $\mathsf{E}_j p$ in B_c^σ. Consequently – assuming p is the only propositional variable – the only formulas weaker than p on which i and j are believed to have expertise are tautologies. Any EC scheme f must therefore have $f_\sigma(i, c, p) \equiv f_\sigma(j, c, p) \equiv \top$. Consequently, EC-selectivity would imply $[B_c^\sigma] = \mathrm{Cn}_0(\top)$. This is a very conservative stance: while there is total consensus for p in case c, p cannot be believed due to disagreement elsewhere. This also conflicts with the "optimistic" attitude described in Example 5.2.1. According to that view we should have $\mathsf{E}_i p \vee \mathsf{E}_j p \in B_c^\sigma$, but this implies $p \in B_c^\sigma$ by **Soundness**, **Containment** and **Closure**. This example already shows that $\mathsf{var\text{-}based\text{-}cond}$, $\mathsf{part\text{-}based\text{-}cond}$ and $\mathsf{excess\text{-}min}$ are not EC-selective.

The core issue is that the expertise of sources is part of the operator's output and is thus uncertain. In order for $\mathsf{E}_i\varphi$ to be believed, i needs to be trusted on φ in *every* maximally plausible world. If there are several such worlds with different assessments of expertise – e.g. if W_1 trusts i but not j, and vice versa in W_2 – then EC-selectivity requires reports to be significantly weakened before expertise can be believed in *all* worlds.

For model-based operators satisfying **Soundness**, this phenomenon can be formalised: a report of φ from source i is expanded by the *join*[16] of the partitions Π_i^W, for $W \in \mathcal{Y}_\sigma$. As the above example shows, this join may be strictly coarser than any of the individual partitions Π_i^W. Formally, for a set of worlds $S \subseteq \mathcal{W}$, write $\Pi_i^S = \bigvee_{W \in S} \Pi_i^W$ for the join of the i-partitions of worlds in S. We first need a preliminary result.

Lemma 5.5.3. *For any model-based operator, $\mathsf{E}_i\varphi \in B_c^\sigma$ iff $\Pi_i^{\mathcal{Y}_\sigma}[\varphi] = \|\varphi\|$.*

Proof. "if": Suppose $\Pi_i^{\mathcal{Y}_\sigma}[\varphi] = \|\varphi\|$. Take $W \in \mathcal{Y}_\sigma$. Then since Π_i^W refines $\Pi_i^{\mathcal{Y}_\sigma}$, we have $\Pi_i^W[\varphi] \subseteq \Pi_i^{\mathcal{Y}_\sigma}[\varphi] = \|\varphi\|$. Since $\|\varphi\| \subseteq \Pi_i^W[\varphi]$ always holds, we have $\Pi_i^W[\varphi] = \|\varphi\|$ and thus $W, c \models \mathsf{E}_i\varphi$. Since $W \in \mathcal{Y}_\sigma$ was arbitrary, this shows $\mathsf{E}_i\varphi \in B_c^\sigma$.

"only if": Suppose $\mathsf{E}_i\varphi \in B_c^\sigma$. We need to show $\Pi_i^{\mathcal{Y}_\sigma}[\varphi] \subseteq \|\varphi\|$. Take $v \in \Pi_i^{\mathcal{Y}_\sigma}[\varphi]$. Let R_i^W be the equivalence relation corresponding to the partition Π_i^W. Then the relation R corresponding to the join $\Pi_i^{\mathcal{Y}_\sigma}$ is the smallest equivalence relation containing each of the R_i^W, which is given explicitly by the transitive closure $R = \left(\bigcup_{W \in \mathcal{Y}_\sigma} R_i^W\right)^+$.

Now, from $v \in \Pi_i^{\mathcal{Y}_\sigma}[\varphi]$ there is some $u \in \|\varphi\|$ such that vRu. By definition of the transitive closure, there are $x_0, \ldots, x_n \in \mathcal{V}$ such that $v = x_0$, $u = x_n$, and for each $0 \leq k < n$, $(x_k, x_{k+1}) \in \bigcup_{W \in \mathcal{Y}_\sigma} R_i^W$. That is, there are $W_0, \ldots, W_{n-1} \in \mathcal{Y}_\sigma$ such that $(x_k, x_{k+1}) \in R_i^{W_k}$. We will show that each x_k lies in $\|\varphi\|$ by backwards induction. For $k = n$, we have $x_n = u \in \|\varphi\|$ by assumption. If $x_{k+1} \in \|\varphi\|$, then $(x_k, x_{k+1}) \in R_i^{W_k}$ gives $x_k \in \Pi_i^{W_k}[x_{k+1}] \subseteq \Pi_i^{W_k}[\varphi]$. By assumption, $\mathsf{E}_i\varphi \in B_c^\sigma$. Since $W_k \in \mathcal{Y}_\sigma$, this means $W_k, c \models \mathsf{E}_i\varphi$ and $\Pi_i^{W_k}[\varphi] = \|\varphi\|$. Hence $x_k \in \|\varphi\|$ as desired. This shows $v = x_0 \in \|\varphi\|$, and we are done. □

Proposition 5.5.2. *If a model-based operator is EC-selective and satisfies **Soundness**, then*

$$\|[B_c^\sigma]\| = \bigcap_{\langle i,\varphi\rangle \in \sigma\restriction c} \Pi_i^{\mathcal{Y}_\sigma}[\varphi]$$

for all $*$-consistent σ and $c \in \mathcal{C}$.

[16] The join of a set of partitions is its least upper bound with respect to the refinement order.

Proof. Let σ be an $*$-consistent sequence and take $c \in \mathcal{C}$.

$\subseteq$: Let $v \in \|[B_c^\sigma]\|$. By Lemma 5.4.1, there is $W \in \mathcal{Y}_\sigma$ such that $v = v_c^W$. Take $\langle i, \varphi\rangle \in \sigma \upharpoonright c$. By **Soundness** (and **Containment**, which holds for all model-based operators) we have $\mathsf{S}_i\varphi \in B_c^\sigma$, so $W, c \models \mathsf{S}_i\varphi$. Consequently $v = v_c^W \in \Pi_i^W[\varphi] \subseteq \Pi_i^{\mathcal{Y}_\sigma}[\varphi]$, where we use the fact that Π_i^W refines $\Pi_i^{\mathcal{Y}_\sigma}$ in the last step.

$\supseteq$: Let $v \in \bigcap_{\langle i,\varphi\rangle \in \sigma \upharpoonright c} \Pi_i^{\mathcal{Y}_\sigma}[\varphi]$. By EC-selectivity there is some expertise-compatible selection scheme f. Write $F_\sigma(i, c, \varphi) = \|f_\sigma(i, c, \varphi)\|$. Then we have

$$\|[B_c^\sigma]\| = \bigcap_{\langle i,\varphi\rangle \in \sigma \upharpoonright c} F_\sigma(i, c, \varphi) \tag{5.5}$$

and $\mathsf{E}_i f_\sigma(i, c, \varphi) \in B_c^\sigma$ for each $\langle i, \varphi\rangle \in \sigma \upharpoonright c$. By Lemma 5.5.3, $\Pi_i^{\mathcal{Y}_\sigma}[F_\sigma(i, c, \varphi)] = F_\sigma(i, c, \varphi)$. We show $v \in \|[B_c^\sigma]\|$ using (5.5). Take $\langle i, \varphi\rangle \in \sigma \upharpoonright c$. By definition of a selection scheme we have $f_\sigma(i, c, \varphi) \in \mathrm{Cn}_0(\varphi)$, so $\|\varphi\| \subseteq F_\sigma(i, c, \varphi)$. Since $v \in \Pi_i^{\mathcal{Y}_\sigma}[\varphi]$ by assumption, we get

$$v \in \Pi_i^{\mathcal{Y}_\sigma}[\varphi] \subseteq \Pi_i^{\mathcal{Y}_\sigma}[F_\sigma(i, c, \varphi)] = F_\sigma(i, c, \varphi)$$

as required. $\square$

Note that Proposition 5.5.2 immediately implies selectivity with respect to any scheme f such that $\|f_\sigma(i, c, \varphi)\| = \Pi_i^{\mathcal{Y}_\sigma}[\varphi]$. Since the right-hand side does not depend on the case c, we get the following corollary.

Corollary 5.5.1. *If a model-based operator is EC-selective and satisfies **Soundness**, then it is case-independent-selective.*

Proposition 5.5.2 also shows that propositional beliefs in case c are determined only by the reports in $\sigma \upharpoonright c$ together with the expertise part of B^σ, via the partitions Π_i^W for $W \in \mathcal{Y}_\sigma$. This property can be expressed syntactically as follows, where for a collection G we write $\mathsf{E}(G)$ for the sub-collection of formulas of the form $\mathsf{E}_i\varphi$.

> **Determination**. For any $*$-consistent σ and $c \in \mathcal{C}$, $[B_c^\sigma] = [\mathrm{Cn}_c(K^\sigma \sqcup \mathsf{E}(B^\sigma))]$.

In other words, **Determination** says that propositional beliefs may be fully recovered by taking (the c-consequences of) the knowledge set K^σ together with just the expertise formulas in B^σ. Surprisingly, **Determination** in fact characterises EC-selectivity, under additional mild assumptions. In what follows, recall that G_{snd}^σ denotes the collection with $(G_{\mathsf{snd}}^\sigma)_c = \{\mathsf{S}_i\varphi \mid \langle i, \varphi\rangle \in \sigma \upharpoonright c\}$.

Theorem 5.5.3. *A model-based operator satisfying **Consistency** and $K^\sigma = \mathrm{Cn}(G_{\mathsf{snd}}^\sigma)$ for all σ is EC-selective if and only if it satisfies **Determination**.*

Proof. Take any model-based operator satisfying **Consistency** and which has $K^\sigma = \mathrm{Cn}(G^\sigma_{\mathsf{snd}})$ for all σ. Note that the latter property implies **Soundness**.

"if": Suppose **Determination** holds. We claim that for any $*$-consistent σ and $c \in \mathcal{C}$,

$$\|[B^\sigma_c]\| = \bigcap_{\langle i,\varphi\rangle \in \sigma\upharpoonright c} \Pi_i^{\mathcal{Y}_\sigma}[\varphi]. \tag{5.6}$$

This implies selectivity upon letting $f_\sigma(i, c, \varphi)$ be any formula with models $\Pi_i^{\mathcal{Y}_\sigma}[\varphi]$. Furthermore it implies EC-selectivity by Lemma 5.5.3.

The left-to-right inclusion of (5.6) follows by an argument identical to that of Proposition 5.5.2 using **Soundness**. It suffices to show the right-to-left inclusion. Take $v \in \bigcap_{\langle i,\varphi\rangle\in\sigma\upharpoonright c} \Pi_i^{\mathcal{Y}_\sigma}[\varphi]$. By **Consistency**, there is some $W_0 \in \mathcal{Y}_\sigma$. Consider W obtained from W_0 by setting its c-valuation to v, and by setting the partition of source i to $\Pi_i^{\mathcal{Y}_\sigma}$:

$$v_d^W = \begin{cases} v^{W_0}, & d \neq c \\ v, & d = c \end{cases},$$

$$\Pi_i^W = \Pi_i^{\mathcal{Y}_\sigma}.$$

Note that since $W_0 \in \mathcal{Y}_\sigma$, $\Pi_i^{W_0}$ refines Π_i^W for each i. We aim to show $W \in \mathrm{mod}(K^\sigma \sqcup \mathsf{E}\,(B^\sigma))$. Recall that, by assumption, $K^\sigma = \mathrm{Cn}(G^\sigma_{\mathsf{snd}})$. It therefore suffices to show that $W \in \mathrm{mod}(G^\sigma_{\mathsf{snd}}) \cap \mathrm{mod}(\mathsf{E}\,(B^\sigma))$. First take $\langle i, d, \varphi\rangle \in \sigma$. By **Soundness** and **Containment** we have $W_0, d \models \mathsf{S}_i\varphi$, i.e. $v_d^{W_0} \in \Pi_i^{W_0}[\varphi]$. If $d \neq c$ then

$$v_d^W = v_d^{W_0} \in \Pi_i^{W_0}[\varphi] \subseteq \Pi_i^W[\varphi],$$

where we use the fact that $\Pi_i^{W_0}$ refines Π_i^W in the last step. Thus $W, d \models \mathsf{S}_i\varphi$ as required. If instead $d = c$, then by our assumption on v,

$$v_c^W = v \in \Pi_i^{\mathcal{Y}_\sigma}[\varphi] = \Pi_i^W[\varphi]$$

so that $W, c \models \mathsf{S}_i\varphi$ as required. This shows $W \in \mathrm{mod}(G^\sigma_{\mathsf{snd}})$. For $W \in \mathrm{mod}(\mathsf{E}\,(B^\sigma))$, take any $\mathsf{E}_i\varphi \in B^\sigma_c$ (note that by **Closure** $\mathsf{E}\,(B^\sigma)$ contains the same formulas in each case, so we may choose c without loss of generality). Then by Lemma 5.5.3, $\Pi_i^{\mathcal{Y}_\sigma}[\varphi] = \|\varphi\|$. By construction of W we evidently have $W, c \models \mathsf{E}_i\varphi$.

This shows $W \in \mathrm{mod}(K^\sigma \sqcup \mathsf{E}\,(B^\sigma))$. Finally, to show $v \in \|[B^\sigma_c]\|$, take any $\psi \in [B^\sigma_c]$. By **Determination**, $\psi \in \mathrm{Cn}_c(K^\sigma \sqcup \mathsf{E}\,(B^\sigma))$. Thus $W, c \models \psi$. But by construction the c-valuation in W is v, so $v \in \|\psi\|$ and we are done.

"only if": Suppose the operator is EC-selective according to some scheme f. To show **Determination**, take any $*$-consistent σ and $c \in \mathcal{C}$. By **Containment** we have $K^\sigma \sqsubseteq B^\sigma$, and clearly $\mathsf{E}\,(B^\sigma) \sqsubseteq B^\sigma$. Consequently $\mathsf{K}^\sigma \sqcup \mathsf{E}\,(B^\sigma) \sqsubseteq B^\sigma$; by monotonicity of Cn and **Closure** we get $\mathrm{Cn}(K^\sigma \sqcup \mathsf{E}\,(B^\sigma)) \sqsubseteq B^\sigma$. This in turn implies $[B^\sigma_c] \supseteq [\mathrm{Cn}_c(K^\sigma \sqcup \mathsf{E}\,(B^\sigma))]$.

For the reverse inclusion, it is sufficient by **Closure** to show

$$\|[\mathrm{Cn}_c(K^\sigma \sqcup \mathsf{E}\,(B^\sigma))]\| \subseteq \|[B^\sigma_c]\|. \tag{5.7}$$

So, take v in the set on the left-hand side. By an argument identical to the proof of Lemma 5.4.1, there is some $W \in \mathrm{mod}(K^\sigma \sqcup \mathsf{E}(B^\sigma))$ such that $v = v_c^W$. Since **Soundness** holds by the assumption that $K^\sigma = \mathrm{Cn}(G^\sigma_{\mathsf{snd}})$, EC-selectivity and Proposition 5.5.2 give $\|[B_c^\sigma]\| = \bigcap_{\langle i,\varphi\rangle \in \sigma \upharpoonright c} \Pi_i^{\mathcal{Y}_\sigma}[\varphi]$.

Take $\langle i, \varphi\rangle \in \sigma \upharpoonright c$. Let ψ be any propositional formula with $\|\psi\| = \Pi_i^{\mathcal{Y}_\sigma}[\varphi]$. Then $\mathsf{E}_i\psi \in B_c^\sigma$, so $W \in \mathrm{mod}(\mathsf{E}(B^\sigma))$ gives $W, c \models \mathsf{E}_i\psi$. Now, **Soundness** and $W \in \mathrm{mod}(K^\sigma)$ also gives $W, c \models \mathsf{S}_i\varphi$, i.e. $v = v_c^W \in \Pi_i^W[\varphi]$. Since $\|\varphi\| \subseteq \Pi_i^{\mathcal{Y}_\sigma}[\varphi] = \|\psi\|$, we get

$$v \in \Pi_i^W[\varphi] \subseteq \Pi_i^W[\psi] = \|\psi\| = \Pi_i^{\mathcal{Y}_\sigma}[\varphi].$$

This shows (5.7) and completes the proof. □

Note that if the basic postulates are given, the condition $K^\sigma = \mathrm{Cn}(G^\sigma_{\mathsf{snd}})$ in Theorem 5.5.3 is equivalent to $K^\emptyset = \mathrm{Cn}(\emptyset)$ by Proposition 5.2.1. In particular, Theorem 5.5.3 applies to our concrete operators $\mathsf{var\text{-}based\text{-}cond}$, $\mathsf{part\text{-}based\text{-}cond}$ and $\mathsf{excess\text{-}min}$. Since we have already seen these operators are *not* EC-selective, we also have that they each fail **Determination**.

The potential problem with EC-selectivity, as expressed by **Determination**, is that it only permits belief formation on the basis of soundness statements together with firmly believed expertise statements in $\mathsf{E}(B^\sigma)$. A natural weaker notion of expertise-compatible selectivity requires not that i is believed to have expertise on $f_\sigma(i, c, \varphi)$, but merely that such expertise is *consistent* with B^σ.

Definition 5.5.4. *A selection scheme f is* weakly expertise-compatible *with an operator $\sigma \mapsto \langle B^\sigma, K^\sigma\rangle$ if for all $*$-consistent σ and $\langle i, c, \varphi\rangle \in \sigma$,*

$$\neg\mathsf{E}_i f_\sigma(i, c, \varphi) \notin B_c^\sigma.$$

Mirroring earlier terminology, say an operator is *weakly EC-selective* if it is selective according to some weakly expertise-compatible scheme. Weak EC-selectivitity overcomes the issues of EC-selectivity highlighted above on the sequence

$$\sigma = (\langle i, c, p\rangle, \langle j, c, p\rangle, \langle i, d, p\rangle, \langle j, d, \neg p\rangle).$$

For example, each of our example operators $\mathsf{var\text{-}based\text{-}cond}$, $\mathsf{part\text{-}based\text{-}cond}$ and $\mathsf{excess\text{-}min}$ are weakly EC-selective for this particular σ according to the selection

$$f_\sigma(i, c, p) = f_\sigma(j, c, p) = p$$
$$f_\sigma(i, d, p) = f_\sigma(j, d, \neg p) = \top.$$

However, this selection is *not* case independent. Questions around the interaction between weak EC-selectivity and case independence, as well as the whether the example operators are weakly EC-selective and/or case-independent-selective in general, are left for future work.

5.6 Related Work

Belief Merging. In the framework of Konieczny and Pino Pérez (2002), a merging operator Δ maps a multiset of propositional formulas $\Psi = \{\varphi_1, \ldots, \varphi_n\}$ and an integrity constraint μ to a formula $\Delta_\mu(\Psi)$. Here φ_i represents the input from source i, the integrity constraint μ represents sure information which must be respected – akin to reports from $*$ in our framework – and $\Delta_\mu(\Psi)$ represents the merged result. Various operators and postulates have been proposed in the literature; see (Konieczny and Pino Pérez 2011) for a review.

Merging can be seen as a special case of our framework, if we impose an upper bound N on the size of the multisets considered as inputs. Indeed, instantiating our framework with $\mathcal{S} = \{1, \ldots, N, *\}$ and a single case $\mathcal{C} = \{c\}$, we can interpret a multiset Ψ and integrity constraint μ as a sequence $\sigma_{\Psi,\mu}$, where

$$\sigma_{\Psi,\mu} = (\langle *, c, \mu\rangle, \langle 1, c, \varphi_1\rangle, \ldots \langle n, c, \varphi_n\rangle).$$

That is, $*$ reports the integrity constraint and each source i reports φ_i.[17] In this way, any operator gives rise to a merging operator – up to logical equivalence – by setting

$$\|\Delta_\mu(\Psi)\| = \|[B_c^{\sigma_{\Psi,\mu}}]\|. \tag{5.8}$$

In fact, for our specific operators $\mathsf{var\text{-}based\text{-}cond}$, $\mathsf{part\text{-}based\text{-}cond}$ and $\mathsf{excess\text{-}min}$, the corresponding merging operators Δ^{vbc}, Δ^{pbc} and Δ^{exm} coincide with well-known *model-based* merging operators.

Definition 5.6.1 (Konieczny and Pino Pérez (2002) and Konieczny and Pino Pérez (2011))**.** *Let $d : \mathcal{V} \times \mathcal{V} \to \mathbb{R}_{\geq 0}$ be a function such that $d(u, v) = d(v, u)$ and $d(u, v) = 0$ iff $u = v$.*[18] *The merging operator $\Delta^{d,\Sigma}$ is defined (up to logical equivalence) by*

$$\|\Delta_\mu^{d,\Sigma}(\Psi)\| = \operatorname*{argmin}_{v \in \|\mu\|} \sum_{i=1}^{n} \min_{u \in \|\varphi_i\|} d(u, v).$$

That is, $\Delta_\mu^{d,\Sigma}(\Psi)$ selects the models of the integrity constraint μ which minimise the sum of the distances to each formula φ_i, with the distance between v and φ_i interpreted as the minimal distance between v and some φ_i model.

Typical distances d include the *Hamming distance* d_H, where $d_H(u, v)$ is the number of propositional variables on which u and v differ, and the *drastic distance* d_D, where $d_D(u, v)$ is 0 if $u = v$ and 1 otherwise. Our operators give rise to model-based merging operators corresponding to these distances.

Theorem 5.6.1. $\Delta^{\mathsf{vbc}} \equiv \Delta^{d_H,\Sigma}$*,* $\Delta^{\mathsf{pbc}} \equiv \Delta^{d_D,\Sigma}$ *and* $\Delta^{\mathsf{exm}} \equiv \Delta^{d_D,\Sigma}$*.*

[17] Since multisets are not ordered, to ensure $\sigma_{\Psi,\mu}$ is well-defined we order the reports $\langle i, c, \varphi_i\rangle$ according to some arbitrary but fixed total order on $\mathcal{L}_0$.

[18] Such d are called *distance functions* in the merging literature, although the triangle inequality $d(u, v) \leq d(u, w) + d(w, v)$ is not required to hold.

The proof can be found in Appendix B.1. It follows that Δ^{vbc}, Δ^{pbc} and Δ^{exm} satisfy the *IC postulates* **(IC0)** – **(IC8)** of Konieczny and Pino Pérez (2002).

Also note that, perhaps surprisingly, part-based-cond and excess-min result in the same merging operator. In some sense this highlights the restrictiveness of purely propositional merging, given that the two operators differ substantially in our general setting (e.g. on **Strong-cond-success** and in Example 5.3.3).[19]

Indeed, we go beyond propositional merging by considering multiple cases and explicitly modelling expertise (and trust, via beliefs about expertise). While it may be possible to model expertise *implicitly* in belief merging (for example, say i is not trusted on ψ if $\Delta_\mu(\Psi) \nvdash \psi$ when $\varphi_i \vdash \psi$), bringing expertise to the object level allows us to express more complex beliefs about expertise, such as $\mathsf{E}_{\mathsf{X}} p \vee \mathsf{E}_{\mathsf{Y}} p$ in Example 5.2.1. It also facilitates postulates which refer directly to expertise, such as the weakenings of **Success** in Section 5.4. Moreover, such beliefs about expertise cannot be recovered from propositional beliefs alone, as demonstrated by the fact that part-based-cond and excess-min differ in general but coincide as merging operators.

However, while more general on a technical level, our problem is more specialised than merging, since we focus specifically on conflicting information due to lack of expertise. Belief merging may be applied more broadly to other types of *information fusion*, e.g. subjective beliefs or goals (Grégoire and Konieczny 2006), where notions of objective expertise do not apply. While our framework *could* be applied in these settings, our postulates and operators may no longer be desirable.

Furthermore, there are further postulates in the belief merging literature which cannot be expressed in our framework due to the fixed-source assumption. For example, consider the *majority* postulate:

Majority. $\exists n \in \mathbb{N} : \ \Delta_\mu(\Psi_1 \sqcup \Psi_2^n) \vdash \Delta_\mu(\Psi_2)$

Clearly **Majority** requires one to consider multisets of unbounded size. A variable-domain approach to our belief change problem – such as the framework for truth discovery in Chapter 2, where the set of sources was given as part of the input – would allow such postulates to be expressed.

Trust and belief revision. Yasser and Ismail (2020) and Yasser and Ismail (2021) study the joint revision of belief and trust in the style of belief revision theory. As with our framework, they consider reports from multiple sources, and set out several postulates to govern the interaction between trust in sources and belief in formulas. Unlike our work, however, they take a more general view of trust – not based on any fixed semantic notion such as expertise –

[19] Note that it is *not* the case that part-based-cond and excess-min output the same propositional beliefs given any input σ; Theorem 5.6.1 only shows this to be the case for σ of the form $\sigma_{\Psi,\mu}$.

wherein sources are assigned degrees of trust on each topic. In this way, their view of trustworthiness is closer to that of truth discovery in Chapter 2.

Formally, they consider totally ordered sets $\mathcal{D}_b$ and $\mathcal{D}_t$ of *belief and trust degrees*, respectively, and a finite collection of *topics* $\mathcal{O}$, where each $T \in \mathcal{O}$ is a set of formulas such that $\bigcup \mathcal{O} = \mathcal{L}$.[20] An *information state* $\mathcal{K}$ then consists of (i) a set of reports of the form $\langle i, \varphi \rangle$; (ii) a partial function assigning degrees of belief in formulas of $\mathcal{L}$; and (iii) a partial function assigning degrees of trust in a source-topic pair.

A revision operator $\ltimes$ then takes an information state $\mathcal{K}$ and a new report $\langle i, \varphi \rangle$ and produces a new information state $\mathcal{K} \ltimes \langle i, \varphi \rangle$. The authors introduce natural notions of *entrenchment*, whereby a formula φ may be more entrenched in a state $\mathcal{K}$ over $\mathcal{K}'$, and similarly where a source i may be more trusted on a topic T in $\mathcal{K}$ over $\mathcal{K}'$. Such entrenchment notions are used to state postulates to constrain $\ltimes$.

This framework generalises ours along several dimensions. Primarily, it allows an operator to consider graded belief and trust via the degree sets $\mathcal{D}_b$ and $\mathcal{D}_t$. In our framework we have, roughly speaking, only three degrees: (i) 1, if $\varphi \in B^\sigma_c$; (ii) -1, if $\neg\varphi \in B^\sigma_c$; and (iii) 0, if $\varphi \notin B^\sigma_c$ and $\neg\varphi \notin B^\sigma_c$. Likewise, we have three degrees for trust by considering membership of $\mathsf{E}_i\varphi$ and $\neg\mathsf{E}_i\varphi$.

Secondly, their notion of trust is vastly more general than ours. This can be seen as a benefit, since the framework can be applied in more settings, or as a drawback, since the generality restricts the extent to which the authors can state postulates which are reasonable in the general case. Indeed, part of our motivation to study trust via expertise was to be able to introduce more specific – and, hopefully, more interesting – postulates and operators.

Our notion of separate cases – representing different instantiations of some propositional domain – can also be represented via an encoding trick. Namely, one can consider the propositional language $\mathcal{L}_0^\mathcal{C}$ formed from propositional variables of the form p_c, for $p \in \mathsf{Prop}$ and $c \in \mathcal{C}$, read as "p holds in case c". A report $\langle i, c, \varphi \rangle$ in our framework becomes $\langle i, \varphi_c \rangle$, where φ_c is obtained from $\varphi \in \mathcal{L}_0$ by replacing each variable p appearing in φ with p_c. To ensure that trust is fixed across cases, one can choose as topics $T_\varphi = \{\varphi_c \mid c \in \mathcal{C}\}$.

Ultimately, our work is complementary and addresses the problem of trust and belief revision from a different angle. Future work could investigate further links and differences between the two frameworks; for example by comparing postulates.

Building trust. In recent work, Hunter (2021) investigates how trust in a source may be determined from its record on past reports. Our work shares a common ancestor via (Booth and Hunter 2018), in which trust is rooted in expertise and the ability of sources to distinguish between states. In (Booth and Hunter 2018), a single partition was used to represent expertise. Here, we consider possibly several partitions Π_i^W, for $W \in \mathcal{Y}_\sigma$.

[20]Note that we have changed the notation compared to the original paper.

Hunter (2021) considers a richer representation of distinguishability, in the form of a *pseudo-ultrametric* d on states for each source; that is, a function $d : \mathcal{V} \times \mathcal{V} \to \mathbb{R}_{\geq 0}$ such that (i) $d(v,v) = 0$; (ii) $d(u,v) = d(v,u)$; and (iii) $d(u,v) \leq \max\{d(u,w), d(w,v)\}$. Here $d(u,v)$ represents the degree to which the source is trusted to be able to distinguish states u and v. Due to the so-called *ultrametric inequality* (iii), the set of balls of radius r – i.e. $\{\{u \mid d(u,v) \leq n\} \mid v \in \mathcal{V}\}$ – forms a partition of $\mathcal{V}$. Consequently, any threshold value r gives rise to a partition. In this sense the pseudo-ultrametric representation generalises partitions.

More importantly, Hunter (2021) considers how to iteratively *revise* a pseudo-ultrametric from a so-called *report history*. Such a history consists of reports of the form $\langle \varphi, j \rangle$, representing the fact that the source had previously reported φ and was either correct (if $j = 1$) or incorrect (if $j = 0$). An algorithm is given to update a pseudo-ultrametric d given a history, and thereby modify the revision agent's perception of the source's expertise on the basis of their past performance.

Report histories can be modelled in our framework by reports from $*$. For example, a negative example $\langle \varphi, 0 \rangle$ from source i can be represented by the sequence $(\langle i, c, \varphi \rangle, \langle *, c, \neg\varphi \rangle)$. However, we take a different view on what to *do* with such histories: our operators select possible partitions representing the source's expertise, whereas the algorithm of Hunter (2021) updates a pseudo-ultrametric representation. Investigating ways in which the two approaches can be combined is an interesting direction for future work.

5.7 Conclusion

Summary. In this chapter we studied a belief change problem – extending the classical AGM framework – in which beliefs about the state of the world in multiple cases, as well as expertise of multiple sources, must be inferred from a sequence of reports. This allowed us to take a fresh look at the interaction between trust (seen as *belief in expertise*) and belief. By inferring the expertise of the sources from the reports, we have generalised some earlier approaches to non-prioritised revision which assume expertise (or reliability, credibility, priority etc) is known up-front (e.g. (Fermé and Hansson 1999; Hansson et al. 2001; Booth and Hunter 2018; Delgrande, Dubois, and Lang 2006)). We went on to propose some concrete belief change operators, and explored their properties through examples, postulates, and a notion of selective revision.

We saw that conditioning operators satisfy some desirable properties, and our concrete instances make useful inferences that go beyond weak-mb. However, we have examples in which intuitively plausible inferences are blocked, and conditioning is largely incompatible with **Strong-cond-success**. Score-based operators, and in particular excess-min, were introduced as a possible way around these limitations.

Limitations and future work. There are many possibilities for future work. Firstly, we have a representation result only for conditioning operators. A characterisation of score-based operators – either the class in general or the specific operator excess-min – remains to be found. This would help to further clarify the differences between conditioning and score-based operators. We have also not considered any computational issues. Determining the complexity of calculating the results of our example operators, and the complexity for conditioning and score-based operators more broadly, is left to future work. Secondly, there is scope for deeper postulate-based analysis. For example, there should be postulates governing how beliefs change in case c in response to reports in case d. We could also consider more postulates relating trust and belief, and compare these postulates with those of Yasser and Ismail (2020). Moreover, there are many weaker version of **Success** which have been considered in the literature (e.g. in (Fermé and Hansson 1999; Hansson et al. 2001; Booth and Hunter 2018)); we should compare these against our **Cond-success** and **Strong-cond-success** in future work.

Finally, as mentioned above, our framework only deals with three levels of trust on a proposition: we can believe $\mathsf{E}_i\varphi$, believe $\neg\mathsf{E}_i\varphi$, or neither. Future work could investigate how to extend our semantics to talk about *graded expertise*, and thereby permit more fine-grained degrees of trust (Hunter 2021; Yasser and Ismail 2020; Delgrande, Dubois, and Lang 2006).

Outlook. Broadly speaking, this chapter addressed *normative* properties of operators, i.e. properties which "reasonable" operators should satisfy. We did not consider any notion of *truthfulness*, i.e. whether or not the belief set B_c^σ is true in the "actual" world. Thus, while the operators introduced here may be *rational* – in that they satisfy the postulates – we cannot say whether they are *truth-tracking*.

The following chapter addresses this gap, by combining our belief change framework with learning-theoretic notions from formal learning theory. In doing so we study truth-tracking; e.g. by investigating the extent to which the truth can be found with non-expert sources and determining which operators are able to track the truth.

— 6 —

Truth-Tracking with Non-Expert Sources

In this chapter we apply the framework of the previous chapter to study *truth-tracking*. Broadly speaking, the goal of truth-tracking is to find the true state of the world given some input which describes it. In our case this involves finding the true state of some propositional domain about which the sources give reports, and finding the extent of the expertise of the sources themselves.

The general problem of truth-tracking has been studied in various forms across many domains. Perhaps the oldest approach goes back to Condorcet (1785), whose celebrated *Jury Theorem* states that a majority vote on a yes/no issue will yield the "correct" answer with probability approaching 1 as the number of voters tends to infinity, provided that each voter is more reliable than random choice. This result has since been generalised in many directions (Grofman, Owen, and Feld 1983). More widely, *epistemic social choice* (Elkind and Slinko 2016) studies aggregation methods (e.g. voting rules) from the point of finding the "correct" result with high probability, where individual votes are seen as noisy approximations. Of particular relevance to our work is truth-tracking in *judgement aggregation* in social choice (Hartmann and Sprenger 2012; Terzopoulou and Endriss 2019), which also takes place in a logical framework. *Belief merging* has close links with judgement aggregation, and generalised jury theorems have been found here too (Everaere, Konieczny, Marquis, et al. 2010).

The problem of *truth discovery* (Y. Li, Gao, et al. 2016), from the crowd-sourcing literature and familiar from Chapter 2, looks at how information from unreliable sources can be aggregated to find the true claims associated with a number of objects, and to find the true reliability level of the sources. Work in this area typically combines empirical results (e.g. how well methods find the truth on test datasets for which true values are known) and theoretical guarantees, and is typically set in a probabilistic framework.

On the other hand, *formal learning theory* (Jain et al. 1999) offers a non-probabilistic view on truth-tracking, stemming from the classical framework of Gold (1967) for identification in the limit. In this paradigm a learner receives an infinite sequence of information step-by-step, such that all true information eventually appears in the sequence. The learner outputs a hypothesis at each step, and aims to stabilise on the correct hypothesis after some finite number of steps. This framework has been combined with belief revision theory (Kelly, Schulte, and Hendricks 1997; Baltag, Gierasimczuk, and Smets 2019) and dynamic epistemic logic (Baltag, Gierasimczuk, Özgün, et al. 2019).

This is the approach we take, and in particular we adapt the truth-tracking setting of Baltag, Gierasimczuk, and Smets (2019). We apply this to the logical framework of Chapter 5, which extends a finitely-generated propositional language with two new notions: that of a source having expertise on a formula, and a formula being sound for a source to report. Consequently we can consider learning both the *ontic* facts of the world, via purely propositional formulas, and the *epistemic* state of the sources, via expertise and soundness formulas.

For the most part, formal learning theory supposes that all information received is true, and that all true information is eventually received.[1] This is not a tenable assumption with non-expert sources: some sources may simply lack the expertise to know whether φ is true or false. In the bulk of this chapter we make a different (and strong) assumption: all and only *sound reports* are received. Thus, sources report everything consistent with their expertise, which necessitates inconsistent reports from non-experts. Inputs of this form should therefore be distinguished from the inputs to belief revision and belief merging methods (Alchourrón, Gärdenfors, and Makinson 1985; Konieczny and Pino Pérez 2002) – also propositional formulas – which represent *beliefs* of the reporting sources. Later we go on to model beliefs of sources directly, and sketch an approach to learning in which sources report all and only that which they believe. Some preliminary results are presented for this alternative model.

The following example – a modification of the motivating example from Chapter 5 – informally illustrates the core concepts and will be returned to throughout the chapter.

Example 6.0.1. *Consider a medical scenario in which patient A is checked for conditions p and q. By examining A, a doctor D has expertise to determine whether A has at least one of p or q, but cannot tell which one(s) without a blood test. A test is only available for p, however, so that the technician T performing the test has expertise on p but not q.*

Supposing A in fact suffers from q but not p, D considers each of $p \wedge q$, $\neg p \wedge q$ and $p \wedge \neg q$ possible, whereas T considers both $\neg p \wedge q$ and $\neg p \wedge \neg q$ possible. Assuming both sources report all *they consider possible, their combined expertise leaves $\neg p \wedge q$ as the only possibility. Intuitively, this means we* can *find the true values of p and q in this case.*

Now consider a patient B who suffers from both conditions. D cannot distinguish A and B, so will provide the same reports, and T considers both $p \wedge q$ and $p \wedge \neg q$ possible. In this case T is more knowledgable than D – since

[1] But see Jain et al. (1999, §8.1), which considers inaccurate data of various kinds, and Baltag, Gierasimczuk, and Smets (2019), which consider erroneous reports provided that all errors are eventually corrected.

they consider fewer situations possible – but we cannot narrow down the true value of q. Thus truth-tracking is only possible for p. The second patient still provides useful information, though, since together with the reports on A, T's lack of expertise tells us all the (in)distinctions between states they are able to make. Namely, T cannot distinguish between $p \land q$ and $p \land \neg q$. Thus we can find the truth about T's expertise.

Contributions. This chapter adapts learning-theoretic notions from formal learning theory – and in particular its intersection with belief revision (Baltag, Gierasimczuk, and Smets 2019) – to handle non-expert information sources. We establish the limits of learning in this setting, and conditions under which one can learn the true facts of the world as well as the true extent of the expertise of the sources. We go on to characterise truth-tracking learning methods in terms of syntactic postulates, and look specifically at the methods already introduced in Chapter 5. Finally, we consider a different model in which the beliefs of sources are represented explicitly, and give some preliminary results in this setting.

This chapter is an extended version of Singleton and Booth (2022c), with new material in Sections 6.5.3 and 6.6.

6.1 Preliminaries

In this section we introduce the logical framework, which is largely the same as in Chapter 5 but with minor differences. Most importantly, we exclude the special source $*$.

Syntax. As before, let Prop be a finite set of propositional variables. Let $\mathcal{L}_0$ denote the propositional language generated from Prop. We use $\mathcal{L}_0$ to model the domain underlying the truth-tracking problem; it describes the "ontic" facts of the world, irrespective of the expertise of the sources. Formulas in $\mathcal{L}_0$ will be denoted by lower-case Greek letters (φ, ψ, etc).

Let $\mathcal{S}$ be a finite set of sources. Here we make an important change to the setup of Chapter 5: we do *not* include the special source $*$. Indeed, having access to a completely reliable source of information would somewhat trivialise the truth-tracking problem.

The language $\mathcal{L}$ extends $\mathcal{L}_0$ with expertise and soundness formulas for each source $i \in \mathcal{S}$, and is defined by the following grammar:

$$\Phi ::= p \mid \Phi \land \Phi \mid \neg\Phi \mid \mathsf{E}_i\varphi \mid \mathsf{S}_i\varphi,$$

for $i \in \mathcal{S}$, $p \in \mathsf{Prop}$ and $\varphi \in \mathcal{L}_0$. Formulas in $\mathcal{L}$ will be denoted by upper-case Greek letters (Φ, Ψ etc). Other logical connectives ($\lor$, $\to$, $\leftrightarrow$) are introduced as abbreviations. As usual, we read $\mathsf{E}_i\varphi$ as "i has expertise on φ", and $\mathsf{S}_i\varphi$ as "φ is sound for i". Note that we again restrict the expertise and soundness

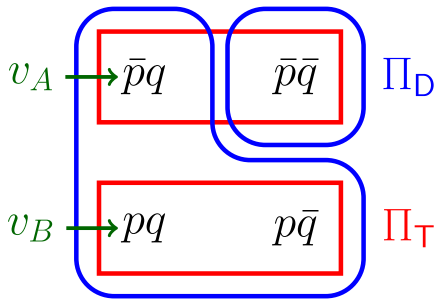


Figure 6.1: Example of a world W, which formalises Example 6.0.1. Here $\mathsf{Prop} = \{p, q\}$*,* $\mathcal{S} = \{\mathsf{D}, \mathsf{T}\}$ *and* $\mathcal{C} = \{A, B\}$*.*

formulas to propositional arguments, and do not consider iterated formulas such as $\mathsf{E}_i\mathsf{S}_j\varphi$.

Semantics. The semantics are identical to those given in Chapter 5; we provide only a brief recap. The set of propositional valuations over Prop is denoted by $\mathcal{V}$. The expertise of a source $i \in \mathcal{S}$ is represented by a partition Π_i of $\mathcal{V}$, which encodes the distinctions between states the source is able to make. We say i has expertise on φ iff i can distinguish all φ states from $\neg\varphi$ states, and φ is sound for i if the "actual" state is indistinguishable from some φ state. $\mathcal{C}$ is a finite set of cases, thought of as independent instantiations of the domain of interest. For example, the cases in Example 6.0.1 are the patients A and B. We consider the expertise of sources to be fixed across all cases. A *world* is a pair $W = \langle \{v_c\}_{c \in \mathcal{C}}, \{\Pi_i\}_{i \in \mathcal{S}} \rangle$, where

- $v_c \in \mathcal{V}$ is the "true" valuation at case $c \in \mathcal{C}$;
- Π_i is a partition of $\mathcal{V}$ representing the "true" expertise of source i.

Let $\mathcal{W}$ denote the set of worlds. Note that $\mathcal{W}$ is finite, since $\mathcal{V}$, $\mathcal{C}$ and $\mathcal{S}$ are. For $\varphi \in \mathcal{L}_0$, write $\|\varphi\| \subseteq \mathcal{V}$ for the models of φ, and write $v \Vdash \varphi$ iff $v \in \|\varphi\|$. The consequences of a set $\Gamma \subseteq \mathcal{L}_0$ is denoted by $\mathrm{Cn}_0(\Gamma)$, and we write $\Gamma \Vdash \varphi$ if $\varphi \in \mathrm{Cn}_0(\Gamma)$. For a partition Π, let $\Pi[v]$ denote the unique cell in Π containing v, and write $\Pi[U] = \bigcup_{v \in U} \Pi[v]$ for $U \subseteq \mathcal{V}$. For brevity, we write $\Pi[\varphi]$ instead of $\Pi[\|\varphi\|]$. We evaluate $\mathcal{L}$ formulas with respect to a world W and a case c as follows:

$$\begin{aligned} W, c \models p &\iff v_c \Vdash p \\ W, c \models \mathsf{E}_i\varphi &\iff \Pi_i[\varphi] = \|\varphi\| \\ W, c \models \mathsf{S}_i\varphi &\iff v_c \in \Pi_i[\varphi] \end{aligned}$$

where the clauses for conjunction and negation are as standard.

Example 6.1.1. *Take W from Fig. 6.1, which formalises Example 6.0.1. Then $W, c \models \mathsf{E}_{\mathsf{D}}(p \vee q)$ for all $c \in \mathcal{C}$, since $\|p \vee q\|$ is a cell in Π_{D}. We also have $W, A \models \neg p \wedge \mathsf{S}_{\mathsf{D}}p$, i.e. patient A does not suffer from condition p, but it is consistent with* D*'s expertise that they do.*

We write $W, c \models \Gamma$, for a set of formulas $\Gamma \subseteq \mathcal{L}$, if $W, c \models \Phi$ for all $\Phi \in \Gamma$. For a set $S \subseteq \mathcal{W}$, we write $S, c \models \Phi$ iff $W, c \models \Phi$ for all $W \in S$.

Reports and methods. A *report* is a triple $\langle i, c, \varphi \rangle$, where $i \in \mathcal{S}$, $c \in \mathcal{C}$ and $\varphi \in \mathcal{L}_0$ with $\varphi \not\equiv \bot$. In most of this chapter, we interpret such triples as source i reporting that φ is *possible* in case c (note that this interpretation was *not* used in Chapter 5). An *input sequence* σ is a finite sequence of reports. Such sequences are inputs to *learning methods*.[2]

Definition 6.1.1. *A* learning method L *maps each input sequence* σ *to a set of worlds* $L(\sigma) \subseteq \mathcal{W}$*, called the* conjecture *of* L *on* σ*.*

We say L *implies* $S \subseteq \mathcal{W}$ on the basis of σ if $L(\sigma) \subseteq S$. L is *consistent* if $L(\sigma) \neq \emptyset$ for all input sequences σ.

This definition is the main point of difference between the framework of the present chapter and of Chapter 5. Firstly, we no longer consider separate belief and knowledge outputs for each sequence σ; the conjecture $L(\sigma)$ should be thought of as analogous to the belief set of an operator in the sense of Definition 5.2.1. Secondly, we take a purely semantic view here by considering the output of a method to be a set of *worlds* instead of a collection of *formulas*. Of course, an operator in the sense of Chapter 5 defines a method by setting $L(\sigma) = \mathrm{mod}(B^\sigma)$, and a method defines the belief set for an operator in the manner of model-based operators (Definition 5.2.2). The difference is therefore in presentation only, and the semantic viewpoint will prove to be more convenient in what follows.

6.2 Truth-Tracking

We adapt the framework for truth-tracking of Baltag, Gierasimczuk, and Smets (2016) and Baltag, Gierasimczuk, and Smets (2019), which finds its roots in formal learning theory. In this framework, a learning method receives increasing initial segments of an infinite sequence – called a *stream* – which enumerates all (and only) the true propositions observable at the "actual" world. Truth-tracking requires the method to eventually find the actual world (or some property thereof), given *any* stream.

As mentioned in the introduction, in our setting we cannot assume the sources themselves report only true propositions. Instead, our streams will enumerate all the *sound* reports. Thus, a stream may include false reports, but such false reports only arise due to lack of expertise of the corresponding source.[3] Moreover, *all* sound reports will eventually arise. Since $\mathsf{S}_i\varphi$ means φ is possible from the point of view of i's expertise, we can view a stream as each

[2] We use the terminology "method" instead of the usual "operator" for consistency with the learning theory literature.

source sharing *all that they consider possible* for each case $c \in \mathcal{C}$. In particular, a non-expert source may report both φ and $\neg\varphi$ for the same case.

Definition 6.2.1. *An infinite sequence of reports ρ is a* stream *for W iff for all i, c, φ:*

$$\langle i, c, \varphi \rangle \in \rho \iff W, c \models \mathsf{S}_i\varphi.$$

We refer to the left-to-right implication as *soundness* of ρ for W, and the right-to-left direction as *completeness*. Note that every world W has some stream: the set $\{\langle i, c, \varphi \rangle \mid W, c \models \mathsf{S}_i\varphi\}$ is countable, so can be indexed by $\mathbb{N}$ to form a stream. For $n \in \mathbb{N}$ we let ρ_n denote the n-th report in ρ, and write $\rho[n]$ for the finite initial segment of ρ of length n.

Example 6.2.1. *Consider W from Fig. 6.1 and case A. From the point of view of D's expertise, the "actual" valuation could be pq, $\bar{p}q$, $p\bar{q}$. Consequently, in a stream for W, D will report p, $\neg p$, q, $\neg q$, $p \lor q$, and so on. A report that D will* not *give is $\neg(p \lor q)$, since D has expertise to know this is false.*

Note that v_A and v_B are indistinguishable to D, so the reports of D in any stream will be the same for both cases. In contrast, T can *distinguish the two cases, and will report $\neg p$ in case A but not in B, and p in case B but not in A.*

A *question* Q is a partition of $\mathcal{W}$. That is, a question is a set of disjoint *answers* $A \in Q$, with each world W appearing in a unique cell $Q[W]$ – the correct answer at W.

Example 6.2.2. *We consider some example questions.*

1. *Any formula $\Phi \in \mathcal{L}$ and case c defines a question $Q_{\Phi,c}$, whose two cells consist of the worlds satisfying Φ, respectively $\neg\Phi$, in case c. Intuitively, this question asks whether Φ is true or false in case c.*

2. *The finest question $Q_\perp = \{\{W\} \mid W \in \mathcal{W}\}$ asks: what is the "actual" world?*

3. *More generally, for any set X and function $f : \mathcal{W} \to X$, the equivalence relation given by $W \simeq_f W'$ iff $f(W) = f(W')$ defines a question Q_f.*

   In this way any data associated with a world gives rise to a question. For example, if $f(W) = \{i \in \mathcal{S} \mid \Pi_i^W[p] = \|p\|\}$ we ask for the set of sources with expertise on p; if $f(W) = |\{c \in \mathcal{C} \mid W, c \models p\}|$ we ask for the number of cases where p holds, etc.

   In fact, all questions are of this form: given Q we may define $f : \mathcal{W} \to Q$ by $f(W) = Q[W]$; then $Q_f = Q$.

[3] Alternatively, we can consider statements of the form "φ is sound for i in case c" as a higher-order "proposition"; a stream then enumerates all true propositions of this kind.

A method solves Q if it eventually implies the correct answer when given any stream.

Definition 6.2.2. *A method L* solves *a question Q if for all worlds W and all streams ρ for W, there is $n \in \mathbb{N}$ such that $L(\rho[m]) \subseteq Q[W]$ for all $m \geq n$. A question Q is* solvable *if there is some consistent method L which solves Q.*

Note that we do not require $W \in L(\rho[m])$. Since we work in a finite framework, solvability can be also expressed in terms of eliminating incorrect worlds.

Proposition 6.2.1. *A method L solves Q if and only if for all W, all streams ρ for W, and all $W' \notin Q[W]$, there is $n_{W'} \in \mathbb{N}$ such that $W' \notin L(\rho[m])$ for all $m \geq n_{W'}$.*

Proof. "if": Taking $n = \max\{n_{W'} \mid W' \notin Q[W]\}$, which exists since $\mathcal{W}$ is finite, $L(\rho[m]) \subseteq Q[W]$ for $m \geq n$.

"only if": Taking n from the definition of L solving Q, we may simply take $n_{W'} = n$ for all $W' \notin Q[W]$. □

6.3 Characterising Solvable Questions

In this section we explore solvability of questions, finding that there is a unique "hardest" question which subsumes all solvable questions. We show this is itself solvable, and thus obtain a precise characterisation of solvability.

Questions are partially ordered by partition refinement: $Q \preceq Q'$ iff each $A' \in Q'$ can be written as a union of answers from Q. Equivalently, $Q[W] \subseteq Q'[W]$ for all W. This can be interpreted as a *difficulty ordering*: if $Q \preceq Q'$ then each answer of Q' is just a disjunction of answers of Q, and thus Q' is *easier* than Q. Naturally, if Q is solvable then so too is any easier question.

Proposition 6.3.1. *If Q is solvable and $Q \preceq Q'$, then Q' is solvable.*

Proof. The method which solves Q also solves Q'. □

Since question solving is based on streams of sound reports, worlds satisfying the same soundness statements cannot be distinguished by any solvable question. To formalise this, define a preorder $\sqsubseteq$ on $\mathcal{W}$ by

$$W \sqsubseteq W' \iff \forall i, c, \varphi : \; (W, c \models \mathsf{S}_i\varphi \implies W', c \models \mathsf{S}_i\varphi).$$

Thus, $W \sqsubseteq W'$ iff any report sound for W is also sound for W'. We denote by $\sqsubset$ and $\approx$ the strict and symmetric parts of $\sqsubseteq$, respectively.[4]

[4]Baltag, Gierasimczuk, and Smets (2016) explore *topological* interpretations of solvability by considering the topology on the set of worlds generated by observable propositions. In our setting, this is the topology generated by sets of the form $\{W \mid W, c \models \mathsf{S}_i\varphi\}$. In this topology, $\sqsubseteq$ is the *specialisation preorder*.

Lemma 6.3.1. *$W \sqsubseteq W'$ if and only if for all $i \in \mathcal{S}$ and $c \in \mathcal{C}$, $\Pi_i^W[v_c^W] \subseteq \Pi_i^{W'}[v_c^{W'}]$.*

Proof. "if": Suppose $W, c \models \mathsf{S}_i\varphi$. Then $v_c^W \in \Pi_i^W[\varphi]$, so there is $u \in \|\varphi\|$ such that $v_c^W \in \Pi_i^W[u]$. Consequently $u \in \Pi_i^W[v_c^W] \subseteq \Pi_i^{W'}[v_c^{W'}]$, which means $v_c^{W'} \in \Pi_i^{W'}[u] \subseteq \Pi_i^{W'}[\varphi]$. Hence $W', c \models \mathsf{S}_i\varphi$. This shows $W \sqsubseteq W'$.

"only if": Let $u \in \Pi_i^W[v_c^W]$. Let φ be any formula with $\|\varphi\| = \{u\}$. Then $W, c \models \mathsf{S}_i\varphi$, so $W \sqsubseteq W'$ gives $W', c \models \mathsf{S}_i\varphi$, i.e. $v_c^{W'} \in \Pi_i^{W'}[u]$, so $u \in \Pi_i^{W'}[v_c^{W'}]$. Hence $\Pi_i^W[v_c^W] \subseteq \Pi_i^{W'}[v_c^{W'}]$. □

Note that $\Pi_i[v_c]$ is the set of valuations indistinguishable from the "actual" valuation in case c, for source i. In light of Lemma 6.3.1, we can interpret $W \sqsubseteq W'$ as saying that all sources are *more knowledgeable* in each case c in world W than in W'. However, $W \sqsubseteq W'$ does not say anything about the partition cells not containing some v_c. Also note that the condition in Lemma 6.3.1 already appeared in the previous chapter, where $\Pi_i^W[v_v^W] \subseteq \Pi_i^{W'}[v_c^{W'}]$ for each $i \in \mathcal{S}$ was called *refinement at* c; this property was used to characterise selective conditioning operators in Proposition 5.5.1.

Proposition 6.3.2. *The following are equivalent.*

1. *W and W' have exactly the same streams.*
2. *$W \approx W'$.*
3. *For all $i \in \mathcal{S}$ and $c \in \mathcal{C}$, $\Pi_i^W[v_c^W] = \Pi_i^{W'}[v_c^{W'}]$.*

Proof. (2) and (3) are easily seen to be equivalent in light of Lemma 6.3.1. To show (1) is equivalent to (2), first suppose W and W' have the same streams, and suppose $W, c \models \mathsf{S}_i\varphi$. Taking an arbitrary stream ρ for W, completeness gives $\langle i, c, \varphi\rangle \in \rho$. But ρ is a stream for W' too, and soundness gives $W', c \models \mathsf{S}_i\varphi$. Hence $W \sqsubseteq W'$. A symmetrical argument shows $W' \sqsubseteq W$.

On the other hand, if $W \approx W'$ then W and W' satisfy exactly the same soundness statements, so it is clear that any sequence ρ is a stream for W iff it is a stream for W'. □

Since it will play a special role throughout, we denote by Q^* the question formed by the equivalence relation $\approx$. Then $Q^*[W]$ is the set of W' with $W \approx W'$. Since no solvable question can distinguish $\approx$-equivalent worlds, we have the following.

Lemma 6.3.2. *If Q is solvable then $Q^* \preceq Q$.*

Proof. Suppose L is a consistent method solving Q. We show $Q^*[W] \subseteq Q[W]$ for all W. Indeed, let $W' \in Q^*[W]$. Then $W' \approx W$. Taking any stream ρ for W, there is n such that $L(\rho[m]) \subseteq Q[W]$ for $m \geq n$. On the other hand ρ is also a stream for W' by Proposition 6.3.2, so there is n' such that $L(\rho[m]) \subseteq Q[W']$ for $m \geq n'$. Setting $m = \max\{n, n'\}$ and using the fact that

L is consistent, we find $\emptyset \subset L(\rho[m]) \subseteq Q[W] \cap Q[W']$. Since Q is a partition, this means $Q[W] = Q[W']$, i.e. $W' \in Q[W]$. $\square$

So, any solvable question is coarser than Q^*. Fortunately, Q^* itself is solvable since we work in a finite framework. For a sequence σ, write $\mathcal{X}^{\mathsf{snd}}_\sigma$ for the set of worlds W such that $W, c \models \mathsf{S}_i\varphi$ for all $\langle i, c, \varphi\rangle \in \sigma$. To solve Q^* it suffices to conjecture the $\sqsubseteq$-minimal worlds in $\mathcal{X}^{\mathsf{snd}}_\sigma$.

Proposition 6.3.3. *Q^* is solvable.*

Proof. Set $L(\sigma) = \min_\sqsubseteq \mathcal{X}^{\mathsf{snd}}_\sigma$ if $\mathcal{X}^{\mathsf{snd}}_\sigma \neq \emptyset$, and $L(\sigma) = \mathcal{W}$ otherwise (where $W \in \min_\sqsubseteq \mathcal{X}^{\mathsf{snd}}_\sigma$ iff $W \in \mathcal{X}^{\mathsf{snd}}_\sigma$ and there is no $W' \in \mathcal{X}^{\mathsf{snd}}_\sigma$ with $W' \sqsubset W$). Note that L is consistent since $\mathcal{W}$ is finite and non-empty. We show that L solves Q^* by Proposition 6.2.1. Take any world W and a stream ρ. First note that, by soundness of ρ, $W \in \mathcal{X}^{\mathsf{snd}}_{\rho[n]}$ for all $n \in \mathbb{N}$, so we are always in the first case in the definition of L.

Take $W' \notin Q^*[W]$. Then $W \not\approx W'$. Consider two cases:

- **Case 1:** $W \not\sqsubseteq W'$. By definition, there are i, c, φ such that $W, c \models \mathsf{S}_i\varphi$ but $W', c \not\models \mathsf{S}_i\varphi$. By completeness of ρ for W, there is n such that $\rho_n = \langle i, c, \varphi\rangle$. Consequently $W' \notin \mathcal{X}^{\mathsf{snd}}_{\rho[m]}$ for all $m \geq n$. Since $L(\rho[m]) \subseteq \mathcal{X}^{\mathsf{snd}}_{\rho[m]}$, we have $W' \notin L(\rho[m])$ as required.
- **Case 2:** $W \sqsubset W'$. Since $W \in \mathcal{X}^{\mathsf{snd}}_{\rho[n]}$ for all n, W' can never be $\sqsubseteq$-minimal. Thus $W' \notin L(\rho[n])$ for all n.

Note that these cases are exhaustive since $W \not\approx W'$. This completes the proof. $\square$

Putting Propositions 6.3.1 and 6.3.3 and Lemma 6.3.2 together we obtain a characterisation of solvable questions.

Theorem 6.3.1. *Q is solvable if and only if $Q^* \preceq Q$.*

Given this result, Q^* is the only question that really matters: any other question is either unsolvable or formed by coarsening Q^*. With this in mind, we make the following definition.

Definition 6.3.1. *A method is* truth-tracking *if it solves Q^*.*

Example 6.3.1. *We refer back to the questions of Example 6.2.2.*

1. *The question $Q_{\varphi,c}$, for any propositional formula $\varphi \in \mathcal{L}_0$, is solvable if and only if either φ is a tautology or a contradiction. To see the "only if" part, consider the contrapositive. For any contingent formula φ, take worlds W_1, W_2 where no source has any expertise (i.e. $\Pi^{W_k}_i = \{\mathcal{V}\}$) but where $v^{W_1}_c \Vdash \varphi$, $v^{W_2}_c \Vdash \neg\varphi$. Then $W_1 \approx W_2$ (e.g. by Proposition 6.3.2) but $W_1 \notin Q_{\varphi,c}[W_2]$.*

   Similarly, $Q_{\mathsf{E}_i\varphi,c}$ is solvable iff either φ is a tautology or contradiction, when $|\mathsf{Prop}| \geq 2$.

2. *The finest question $Q_\perp$ is not solvable, since there are always distinct W, W' with $W \approx W'$.*

3. *In general, Q_f is solvable iff $W \approx W'$ implies $f(W) = f(W')$, i.e. iff f takes a unique value on each equivalence class of $\approx$.*

6.4 What Information can be Learned?

Solving a question Q has a *global* character: we must find the correct answer $Q[W]$ starting from *any* world W. As we saw in Example 6.3.1, this rules out the possibility of solving many interesting questions due to the presence of "abnormal" worlds (e.g. those in which no sources have any expertise). In this section we take a more fine-grained approach by looking *locally*: given some *particular* world W, what can we learn about W via truth-tracking methods? Concretely, what properties of W are uniquely defined across $Q^*[W]$? For instance, in Example 6.0.1 we took this local perspective and argued that in the particular world W modelling the medial scenario, it is possible to find the true value of q for patient A but not for B. The results and examples of this section will formalise this informal argument.

In general, the extent to which one can learn depends on W. If no sources have expertise then source partitions are uniquely defined (since *all* consistent formulas are sound, and only the trivial partitions have this property), but any combination of valuations is possible. On the other hand if all sources have total expertise then valuations are uniquely defined, but there may not be enough cases to uniquely identify the source partitions. Of particular interest is the case where $Q^*[W]$ contains only W; starting in such a world, truth-tracking methods are able to find the true world exactly.

In what follows, say $S \subseteq \mathcal{W}$ *decides* Φ *in case* c iff either $S, c \models \Phi$ or $S, c \models \neg\Phi$. That is, the truth value of Φ in case c is unambiguously defined across S. If Φ does not depend on the case (e.g. if $\Phi = \mathsf{E}_i\varphi$) we simply say S decides Φ.

6.4.1 Valuations

We start by considering when $Q^*[W]$ decides a propositional formula φ in case c, i.e. when truth-tracking methods are guaranteed to successfully determine whether or not φ holds in the "actual" world. This leads to a precise characterisation of when $Q^*[W]$ contains a *unique* valuation in case c, so that v_c^W can be found exactly.

We need a notion of *group expertise*. For $\mathcal{S}' \subseteq \mathcal{S}$ and $\Gamma \subseteq \mathcal{L}_0$, write $W \models \mathsf{E}_{\mathcal{S}'}\Gamma$ if for each $\psi \in \Gamma$ there is $i \in \mathcal{S}'$ such that $W \models \mathsf{E}_i\psi$. Then the group $\mathcal{S}'$ have expertise on Γ in a collective sense, even if no single source has expertise on *all* formulas in Γ.[5] We have that φ is decided if $\mathcal{S}$ has group expertise on a set of true formulas $\Gamma \subseteq \mathcal{L}_0$ such that either $\Gamma \Vdash \varphi$ or $\Gamma \Vdash \neg\varphi$.

Theorem 6.4.1. *$Q^*[W]$ decides $\varphi \in \mathcal{L}_0$ in case c if and only if there is $\Gamma \subseteq \mathcal{L}_0$ such that (i) $W, c \models \Gamma$; (ii) $W \models \mathsf{E}_{\mathcal{S}}\Gamma$; and (iii) either $\Gamma \Vdash \varphi$ or $\Gamma \Vdash \neg\varphi$.*

$Q^*[W]$ decides *all* propositional formulas – and thus determines the c-valuation v_c^W exactly – iff $\mathcal{S}$ has group expertise on a maximally consistent set of true formulas. For $S \subseteq \mathcal{W}$ and $c \in \mathcal{C}$, write $\mathcal{V}_c^S = \{v_c^W \mid W \in S\}$ for the c-valuations appearing in S.

Theorem 6.4.2. *The following are equivalent.*

1. $\mathcal{V}_c^{Q^*[W]} = \{v_c^W\}$.

2. *$Q^*[W]$ decides φ in case c, for all $\varphi \in \mathcal{L}_0$.*

3. *There is $\Gamma \subseteq \mathcal{L}_0$ such that (i) $W, c \models \Gamma$; (ii) $W \models \mathsf{E}_{\mathcal{S}}\Gamma$; and (iii) $\mathrm{Cn}_0(\Gamma)$ is a maximally consistent set.*

We illustrate Theorem 6.4.2 with an example.

Example 6.4.1. *Consider W from Fig. 6.1. Then one can show $\mathcal{V}_A^{Q^*[W]} = \{\bar{p}q\} = \{v_A^W\}$, and $\mathcal{V}_B^{Q^*[W]} = \{pq, p\bar{q}\} \neq \{v_B^W\}$. That is, W's A valuation is uniquely determined by truth-tracking methods, but its B valuation is not: there is some world $W' \approx W$ whose B-valuation differs from W's. This matches the informal reasoning in Example 6.0.1, in which patient A could be successfully diagnosed on both p and q but B could not.*

Formally, take $\Gamma = \{p \vee q, \neg p\}$. Then $W, A \models \Gamma$, $W \models \mathsf{E}_{\mathcal{S}}\Gamma$ (since D has expertise on $p \vee q$ and T has expertise on $\neg p$), and $\mathrm{Cn}_0(\Gamma) = \mathrm{Cn}_0(\neg p \wedge q)$, which is maximally consistent. This example shows how the expertise of multiple sources can be combined to find valuations uniquely, but that this is not necessarily possible in all cases.

The remainder of this section proves Theorems 6.4.1 and 6.4.2.

Lemma 6.4.1. *For $W \approx W'$, $i \in \mathcal{S}$ and $\varphi \in \mathcal{L}_0$,*

$$W, c \models \varphi \wedge \mathsf{E}_i\varphi \implies W', c \models \varphi.$$

Proof. From $W, c \models \varphi$ we have $v_c^W \in \|\varphi\|$, so $\Pi_i^W[v_c^W] \subseteq \Pi_i^W[\varphi]$. But $W, c \models \mathsf{E}_i\varphi$ means $\Pi_i^W[\varphi] = \|\varphi\|$, so in fact $\Pi_i^W[v_c^W] \subseteq \|\varphi\|$. Now using $W \approx W'$, we find $v_c^{W'} \in \Pi_i^{W'}[v_c^{W'}] = \Pi_i^W[v_c^W] \subseteq \|\varphi\|$. Hence $W', c \models \varphi$. □

[5] In contrast to the notions of collective expertise introduced in Chapter 4, here we refer to joint expertise on a *set* of formulas. If one considers a weaker form of distributed expertise in which expertise collections are combined as in Section 4.5.2 but not closed under intersections and unions – e.g. if sources cannot communicate directly with one another – then $\mathsf{E}_{\mathcal{S}'}\Gamma$ corresponds to joint expertise on each $\varphi \in \Gamma$.

Lemma 6.4.2. $\mathcal{V}_c^{Q^*[W]} = \bigcap_{i\in\mathcal{S}} \Pi_i^W[v_c^W]$.

Proof. "$\subseteq$": Suppose $u \in \mathcal{V}_c^{Q^*[W]}$. Then there is $W' \approx W$ such that $u = v_c^{W'}$. Let $i \in \mathcal{S}$. Then $u \in \Pi_i^{W'}[v_c^{W'}] = \Pi_i^W[v_c^W]$ by Proposition 6.3.2, as required.

"$\supseteq$": Suppose $u \in \bigcap_{i\in\mathcal{S}} \Pi_i^W[v_c^W]$. Let W' be the world obtained from W by setting the c-valuation to u, keeping partitions and other valuations the same. We need to show $W' \approx W$. We do so via Proposition 6.3.2, by showing condition (3). Take any $i \in \mathcal{S}$ and $d \in \mathcal{C}$. If $d \neq c$ then $v_d^{W'} = v_d^W$; since partitions are the same in W' as in W we get $\Pi_i^W[v_d^W] = \Pi_i^{W'}[v_d^{W'}]$. For $c = d$, note $\Pi_i^{W'}[v_c^{W'}] = \Pi_i^W[u]$. By assumption $u \in \Pi_i^W[v_c^W]$, so $\Pi_i^W[u] = \Pi_i^W[v_c^W]$. Hence $\Pi_i^{W'}[v_c^{W'}] = \Pi_i^W[v_c^W]$ as required. $\square$

Proof of Theorem 6.4.1. "if": Take $W' \in Q^*[W]$. Note that since $W, c \models \Gamma$ and $W, c \models \mathsf{E}_{\mathcal{S}}\Gamma$, we may apply Lemma 6.4.1 to each formula in Γ in turn to find $W', c \models \Gamma$. Now, if $W, c \models \varphi$ then we must have $\Gamma \Vdash \varphi$, so $W', c \models \varphi$ too. Otherwise $W, c \not\models \varphi$, so we must have $\Gamma \Vdash \neg\varphi$ and $W', c \not\models \varphi$. This shows $W', c \models \varphi$ if and only if $W, c \models \varphi$. Since $W' \in Q^*[W]$ was arbitrary, $Q^*[W]$ decides φ in case c.

"only if": Suppose $Q^*[W]$ decides φ in case c. For each $i \in \mathcal{S}$, take some $\psi_i \in \mathcal{L}_0$ such that $\|\psi_i\| = \Pi_i^W[v_c^W]$. Then $W \models \mathsf{E}_i\psi_i$. Set $\Gamma = \{\psi_i\}_{i\in\mathcal{S}}$. Clearly $W, c \models \Gamma$ and $W \models \mathsf{E}_{\mathcal{S}}\Gamma$. Now, take any $u \in \|\Gamma\|$. By Lemma 6.4.2, $\|\Gamma\| = \bigcap_{i\in\mathcal{S}} \Pi_i^W[v_c^W] = \mathcal{V}_c^{Q^*[W]}$. Hence there is some $W' \in Q^*[W]$ such that $u = v_c^{W'}$. But $Q^*[W]$ decides φ in case c, so $W', c \models \varphi$ iff $W, c \models \varphi$. Thus $u \Vdash \varphi$ iff $W, c \models \varphi$. Since $u \in \|\Gamma\|$ was arbitrary, we have $\Gamma \Vdash \varphi$ if $W, c \models \varphi$, and $\Gamma \Vdash \neg\varphi$ otherwise. $\square$

Proof of Theorem 6.4.2. (1) implies (2): If $W' \in Q^*[W]$ then W and W' share the same c-valuation by (1), so clearly $W, c \models \varphi$ iff $W', c \models \varphi$, for any φ. Hence $Q^*[W]$ decides φ in case c.

(2) implies (1): Clearly $v_c^W \in \mathcal{V}_c^{Q^*[W]}$. Suppose $u \in \mathcal{V}_c^{Q^*[W]}$. Then there is $W' \in Q^*[W]$ such that $u = v_c^{W'}$. Let $p \in \mathsf{Prop}$. Since $W, W' \in Q^*[W]$ and $Q^*[W]$ decides p in case c, we have $u \Vdash p$ iff $v_c^W \Vdash p$. Since p was arbitrary, $u = v_c^W$.

(2) implies (3): Applying Theorem 6.4.1 to each $\varphi \in \mathcal{L}_0$, there is a set $\Gamma_\varphi \subseteq \mathcal{L}_0$ such that $W, c \models \Gamma_\varphi$, $W \models \mathsf{E}_{\mathcal{S}}\Gamma_\varphi$, and either $\Gamma_\varphi \Vdash \varphi$ or $\Gamma_\varphi \Vdash \neg\varphi$. Set $\Gamma = \bigcup_{\varphi\in\mathcal{L}_0} \Gamma_\varphi$. Clearly $W, c \models \Gamma$ – so Γ is consistent – and $W \models \mathsf{E}_{\mathcal{S}}\Gamma$. To show $\mathrm{Cn}_0(\Gamma)$ is *maximally* consistent, suppose $\varphi \notin \mathrm{Cn}_0(\Gamma)$. From monotonicity of classical consequence and $\Gamma_\varphi \subseteq \Gamma$, we get $\varphi \notin \mathrm{Cn}_0(\Gamma_\varphi)$. Hence $\Gamma_\varphi \Vdash \neg\varphi$, and $\Gamma \Vdash \neg\varphi$ too. This means $\mathrm{Cn}_0(\Gamma) \cup \{\varphi\}$ is inconsistent, and we are done.

(3) implies (2): Take $\varphi \in \mathcal{L}_0$. Then we may apply Theorem 6.4.1 with Γ from (3) – noting that the maximal consistency property ensure either $\Gamma \Vdash \varphi$ or $\Gamma \models \neg\varphi$ – to see that $Q^*[W]$ decides φ in case c. $\square$

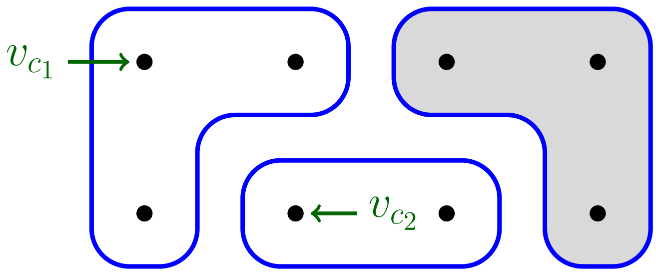


Figure 6.2: World W from Example 6.4.2. Valuations are unlabelled for brevity.

6.4.2 Source Partitions

We now apply the analysis of the previous section to the set of source partitions $\{\Pi_i^W\}_{i \in \mathcal{S}}$ in order to determine conditions under which the true expertise of i can be found by truth-tracking methods. For $S \subseteq \mathcal{W}$ and $i \in \mathcal{S}$, write $\mathcal{P}_i^S = \{\Pi_i^W \mid S \in W\}$ for the i-partitions appearing in S. When $S = Q^*[W]$, these are exactly those partitions which agree with Π_i^W at each valuation v_c^W.

Lemma 6.4.3. $\Pi \in \mathcal{P}_i^{Q^*[W]}$ *if and only if* $\{\Pi_i^W[v_c^W]\}_{c \in \mathcal{C}} \subseteq \Pi$.

Proof. "if": Suppose $\{\Pi_i^W[v_c^W]\}_{c \in \mathcal{C}} \subseteq \Pi$. Let W' be obtained from W by setting i's partition to Π, keeping valuations and other source partitions the same. We claim $W' \approx W$. Indeed, take any $j \in \mathcal{S}$ and $c \in \mathcal{C}$. If $j \neq i$ then $\Pi_j^{W'} = \Pi_i^W$; since valuations are the same we get $\Pi_j^W[v_c^W] = \Pi_j^{W'}[v_c^{W'}]$. For $j = i$, note that since $\Pi_i^W[v_c^W] \in \Pi$ by assumption, and $v_c^W \in \Pi_i^W[v_c^W]$, we have $\Pi[v_c^W] = \Pi_i^W[v_c^W]$. By construction of W', this means $\Pi_i^W[v_c^W] = \Pi[v_c^{W'}] = \Pi_i^{W'}[v_c^{W'}]$. By Proposition 6.3.2, $W' \approx W$. Hence $\Pi \in \mathcal{P}_i^{Q^*[W]}$.

"only if": This is clear from Proposition 6.3.2. □

Example 6.4.2. *Suppose* $|\mathsf{Prop}| = 3$, $\mathcal{C} = \{c_1, c_2\}$ *and* $i \in \mathcal{S}$. *Consider a world W whose i-partition is shown in Fig. 6.2. By Lemma 6.4.3, a partition Π appears as $\Pi_i^{W'}$ for some $W' \approx W$ if and only if it contains the leftmost and bottommost sets. Any such Π consists of these cells together with a partition of the shaded area. Since there are 5 possible partitions of a 3-element set, it follows that* $|\mathcal{P}_i^{Q^*[W]}| = 5$.

Example 6.4.2 hints that if the cells containing the valuations v_c^W cover the whole space of valuations $\mathcal{V}$, or just omit a single valuation, then i's partition is uniquely defined in $Q^*[W]$. That is, truth-tracking methods can determine the full extent of i's expertise if the "actual" world is W. Indeed, we have the following analogue of Theorem 6.4.2 for partitions.

Theorem 6.4.3. *The following are equivalent.*

1. $\mathcal{P}_i^{Q^*[W]} = \{\Pi_i^W\}$.

2. $Q^*[W]$ *decides* $\mathsf{E}_i\varphi$ *for all* $\varphi \in \mathcal{L}_0$.

3. $|\mathcal{V} \setminus R| \leq 1$*, where* $R = \bigcup_{c \in \mathcal{C}} \Pi_i^W[v_c^W]$*.*

Note that $R = \bigcup_{c \in \mathcal{C}} \Pi_i^W[v_c^W]$ is the set of valuations indistinguishable from the actual state at some case c. Theorem 6.4.3 (3) says this set needs to essentially cover the whole space $\mathcal{V}$, omitting at most a single point. In this sense, it is easier to find Π_i^W uniquely when i has *less expertise*, since the cells $\Pi_i^W[v_c^W]$ will be larger. In the extreme case where i has total expertise, i.e. $\Pi_i^W = \{\{v\} \mid v \in \mathcal{V}\}$, we need at least $2^{|\mathsf{Prop}|} - 1$ cases with distinct valuations in order to find Π_i^W exactly.

Example 6.4.3. *In Example 6.4.2 we have already seen an example of a world* W *for which* $\mathcal{P}_i^{Q^*[W]}$ *does not contain a unique partition. For a positive example, consider the world* W *from Fig. 6.1. Then* $\mathcal{V} \setminus R_{\mathsf{D}} = \{\bar{p}\bar{q}\}$ *and* $\mathcal{V} \setminus R_{\mathsf{T}} = \emptyset$*, so both the partitions of* D *and* T *can be found uniquely by truth-tracking methods.*

The remainder of this section proves Theorem 6.4.3.

Lemma 6.4.4. *Let* $i \in \mathcal{S}$ *and* $U \subseteq \mathcal{V}$*. Then* $U \subseteq \bigcup_{c \in \mathcal{C}} \Pi_i^W[v_c^W]$ *and* $W \approx W'$ *implies* $\Pi_i^W[U] = \Pi_i^{W'}[U]$*.*

Proof. It suffices to show that for all $u \in U$ we have $\Pi_i^W[u] = \Pi_i^{W'}[u]$, since by definition $\Pi[U] = \bigcup_{u \in U} \Pi[u]$. Let $u \in U$. Then there is $c \in \mathcal{C}$ such that $u \in \Pi_i^W[v_c^W]$. Hence $\Pi_i^W[u] = \Pi_i^W[v_c^W]$. But since $W \approx W'$, $\Pi_i^W[v_c^W] = \Pi_i^{W'}[v_c^{W'}]$. This means $u \in \Pi_i^{W'}[v_c^{W'}]$, so $\Pi_i^{W'}[u] = \Pi_i^{W'}[v_c^{W'}] = \Pi_i^W[v_c^W] = \Pi_i^W[u]$, as required, □

Lemma 6.4.5. $Q^*[W]$ *decides* $\mathsf{E}_i\varphi$ *if and only if, writing* $R = \bigcup_{c \in \mathcal{C}} \Pi_i^W[v_c^W]$*, either (i)* $\|\varphi\| \subseteq R$*; (ii)* $\|\neg\varphi\| \subseteq R$*; or (iii) there is some* $c \in \mathcal{C}$ *such that* $\Pi_i^W[v_c^W]$ *intersects with both* $\|\varphi\|$ *and* $\|\neg\varphi\|$*.*

Proof. "if": First suppose (i) holds. Take $W' \in Q^*[W]$. From $\|\varphi\| \subseteq R$, $W \approx W'$ and Lemma 6.4.4 we get $\Pi_i^W[\varphi] = \Pi_i^{W'}[\varphi]$. Consequently, $W' \models \mathsf{E}_i\varphi$ iff $W \models \mathsf{E}_i\varphi$. Since W' was arbitrary, either all worlds in $Q^*[W]$ satisfy $\mathsf{E}_i\varphi$, or all do not. Hence $Q^*[W]$ decides $\mathsf{E}_i\varphi$.

If (ii) holds, a similar argument shows that $Q^*[W]$ decides $\mathsf{E}_i\neg\varphi$. But it is easily checked that $\mathsf{E}_i\varphi \equiv \mathsf{E}_i\neg\varphi$, so $Q^*[W]$ also decides $\mathsf{E}_i\varphi$.

Finally, suppose (iii) holds. Then there is $c \in \mathcal{C}$ and $u \in \|\varphi\|$, $v \in \|\neg\varphi\|$ such that $u, v \in \Pi_i^W[v_c^W]$. We claim $Q^*[W] \models \neg\mathsf{E}_i\varphi$. Indeed, take $W' \in Q^*[W]$. Then $\Pi_i^W[v_c^W] = \Pi_i^{W'}[v_c^{W'}]$, so $u, v \in \Pi_i^{W'}[v_c^{W'}]$. In particular, u and v differ on φ but are contained in the same cell in $\Pi_i^{W'}$. Hence $W' \models \neg\mathsf{E}_i\varphi$.

"only if": We show the contrapositive. Suppose none of (i), (ii), (iii) hold. Then there is $u \in \|\varphi\| \setminus R$ and $v \in \|\neg\varphi\| \setminus R$. Let us define two worlds W_1, W_2 from W by modifying i's partition:

$$\Pi_i^{W_1} = \{\Pi_i^W[v_c^W]\}_{c \in \mathcal{C}} \cup \{\mathcal{V} \setminus R\},$$
$$\Pi_i^{W_2} = \{\Pi_i^W[v_c^W]\}_{c \in \mathcal{C}} \cup \{\{w\} \mid w \in \mathcal{V} \setminus R\}.$$

Then $W_1, W_2 \in Q^*[W]$ by Lemma 6.4.3. We claim that $W_1 \models \neg \mathsf{E}_i\varphi$ but $W_2 \models \mathsf{E}_i\varphi$, which will show $Q^*[W]$ does not decide $\mathsf{E}_i\varphi$.

First, note that since $u, v \notin R$, we have $\Pi_i^{W_1}[u] = \Pi_i^{W_1}[v] = \mathcal{V} \setminus R$. Since u and v differ on φ but share the same partition cell, $W_1 \models \neg\mathsf{E}_i\varphi$.

To show $W_2 \models \mathsf{E}_i\varphi$, take $w \in \|\varphi\|$. If $w \notin R$ then $\Pi_i^{W_2}[w] = \{w\} \subseteq \|\varphi\|$. Otherwise there is $c \in \mathcal{C}$ such that $w \in \Pi_i^W[v_c^W]$. Thus $\Pi_i^W[v_c^W]$ intersects with $\|\varphi\|$. Since (iii) does not hold, this in fact implies $\Pi_i^W[v_c^W] \subseteq \|\varphi\|$, and consequently $\Pi_i^{W_2}[w] = \Pi_i^W[v_c^W] \subseteq \|\varphi\|$. Since $w \in \|\varphi\|$ was arbitrary, we have shown $\Pi_i^{W_2}[\varphi] = \bigcup_{w \in \|\varphi\|} \Pi_i^{W_2}[w] \subseteq \|\varphi\|$. Since the reverse inclusion always holds, this shows $W_2 \models \mathsf{E}_i\varphi$, and we are done. $\square$

Proof of Theorem 6.4.3. The implication (1) to (2) is clear since if $W' \in Q^*[W]$ then $\Pi_i^{W'} = \Pi_i^W$ by (1), so $W' \models \mathsf{E}_i\varphi$ iff $W \models \mathsf{E}_i\varphi$, and thus $Q^*[W]$ decides $\mathsf{E}_i\varphi$.

To show (2) implies (3) we show the contrapositive. Suppose $|\mathcal{V} \setminus R| > 1$. Then there are distinct $u, v \in \mathcal{V} \setminus R$. Let φ be any propositional formula with $\|\varphi\| = \{u\}$. We show by Lemma 6.4.5 that $Q^*[W]$ does not decide $\mathsf{E}_i\varphi$. Indeed, all three conditions fail: $\|\varphi\| \not\subseteq R$ (since $u \notin R$), $\|\neg\varphi\| \not\subseteq R$ (since $v \in \|\neg\varphi\| \setminus R$) and no $\Pi_i^W[v_c^W]$ intersects with $\|\varphi\|$ (otherwise $u \in \Pi_i^W[v_c^W] \subseteq R$).

Finally, for (3) implies (1) we also show the contrapositive. Suppose there is $\Pi \in \mathcal{P}_i^{Q^*[W]} \setminus \{\Pi_i^W\}$. Write $\mathcal{R} = \{\Pi_i^W[v_c^W]\}_{c \in \mathcal{C}}$, so that $\mathcal{R}$ is a partition of R. By Lemma 6.4.3, $\mathcal{R} \subseteq \Pi$. Note that $\mathcal{R} \subseteq \Pi_i^W$ too. Since $\Pi \neq \Pi_i^W$, we in fact have $\mathcal{R} \subset \Pi$ and $\mathcal{R} \subset \Pi_i^W$. Hence $\Pi \setminus \mathcal{R}$ and $\Pi_i^W \setminus \mathcal{R}$ are distinct partitions of $\mathcal{V} \setminus R$. Since a one-element set has a unique partition, $\mathcal{V} \setminus R$ must contain at least two elements. $\square$

6.4.3 Learning the Actual World Exactly

Putting Theorems 6.4.2 and 6.4.3, we obtain a precise characterisation of when W can be found *exactly* by truth-tracking methods, i.e when $Q^*[W] = \{W\}$.

Corollary 6.4.1. $Q^*[W] = \{W\}$ *if and only if*

1. *There is a collection* $\{\Gamma_c\}_{c \in \mathcal{C}} \subseteq \mathcal{L}_0^{\mathcal{C}}$ *such that for each* c*, (i)* $W, c \models \Gamma_c$*; (ii)* $W \models \mathsf{E}_{\mathcal{S}}\Gamma_c$*; (iii)* $\mathrm{Cn}_0(\Gamma_c)$ *is maximally consistent; and*

2. *For each each* $i \in \mathcal{S}$*,* $|\mathcal{V} \setminus \bigcup_{c \in \mathcal{C}} \Pi_i^W[v_c^W]| \leq 1$*.*

6.5 Truth-Tracking Methods

So far we have focussed on solvable questions, and the extent to which they reveal information about the actual world. We now turn to the methods which solve them. We give a general characterisation of truth-tracking methods under mild assumptions, before discussing the family of conditioning and score-based methods from Chapter 5.

6.5.1 A General Characterisation

For sequences σ, δ, write $\sigma \equiv \delta$ iff δ is obtained from σ by replacing each report $\langle i, c, \varphi \rangle$ with $\langle i, c, \psi \rangle$, for some $\psi \equiv \varphi$. For $k \in \mathbb{N}$, let σ^k denote the k-fold repetition of σ. Consider the following properties which may hold of a learning method L.

> **Equivalence**. If $\sigma \equiv \delta$ then $L(\sigma) = L(\delta)$.
>
> **Repetition**. $L(\sigma^k) = L(\sigma)$.
>
> **Soundness**. $L(\sigma) \subseteq \mathcal{X}_\sigma^{\mathsf{snd}}$.

Equivalence says that L should not care about the syntactic form of the input. **Repetition** says that the output from L should not change if each source repeats their reports k times. **Soundness** says that all reports in σ are conjectured to be sound, and is the analogue of the same postulate in Chapter 5.

For methods satisfying these properties, we have a precise characterisation of truth-tracking, i.e. necessary and sufficient conditions for L to solve Q^*. First, some new notation is required. Write $\delta \preceq \sigma$ iff for each $\langle i, c, \varphi \rangle \in \delta$ there is $\psi \equiv \varphi$ such that $\langle i, c, \psi \rangle \in \sigma$. That is, σ contains everything in δ, up to logical equivalence. Set

$$T_\sigma = \mathcal{X}_\sigma^{\mathsf{snd}} \setminus \bigcup \left\{ \mathcal{X}_\delta^{\mathsf{snd}} \mid \delta \not\preceq \sigma \right\} \subseteq \mathcal{W}.$$

Then $W \in T_\sigma$ iff σ is sound for W and any δ sound for W has $\delta \preceq \sigma$. In this sense σ contains *all* soundness statements for W – up to equivalence – so can be seen as a finite version of a stream. Let us call σ a *pseudo-stream* for W whenever $W \in T_\sigma$. The truth-tracking characterisation uses the following postulate.

> **Credulity**. If $T_\sigma, c \not\models \mathsf{S}_i\varphi$ then $L(\sigma), c \models \neg\mathsf{S}_i\varphi$.

Theorem 6.5.1. *A method L satisfying **Equivalence**, **Repetition** and **Soundness** is truth-tracking if and only if it satisfies **Credulity**.*

Before the proof, we comment on our interpretation of **Credulity**. It says that whenever $\neg\mathsf{S}_i\varphi$ is consistent with T_σ – those W for which σ is a pseudo-stream – $L(\sigma)$ should imply $\neg\mathsf{S}_i\varphi$. Since the number of sound statements *decreases* with increasing expertise, this is a principle of *maximal trust*: we should believe i has the expertise to rule out φ in case c, whenever this is consistent with T_σ. That is, some amount of *credulity* is required to find the truth. Our assumption that learning methods receive complete streams ensures that, if a source in fact lacks this expertise, they will eventually report φ and this belief can be be retracted. A stronger version of **Credulity** spells this out explicitly in terms of expertise:

$$\text{If } T_\sigma, c \not\models \neg\mathsf{E}_i\varphi \text{ then } L(\sigma), c \models \mathsf{E}_i\varphi. \tag{6.1}$$

We show below that (6.1) implies **Credulity** in the presence of **Soundness**, and is thus a sufficient condition for truth-tracking (when also taken with **Equivalence** and **Repetition**).[6]

Theorem 6.5.1 also shows truth-tracking cannot be performed *deductively*: the method $L(\sigma) = \mathcal{X}_\sigma^{\mathsf{snd}}$ – which does not go beyond the mere information that each report is sound, and corresponds to the weak model-based operator from Definition 5.2.3 – fails **Credulity**.[7] Some amount of or *non-monotonic* reasoning, as captured by **Credulity**, is necessary.

The rest of this section works towards the proof of Theorem 6.5.1. We collect some useful properties of pseudo-streams. First, pseudo-streams provide a way of accessing Q^* via a finite sequence: T_σ is a cell in Q^* whenever it is non-empty.

Lemma 6.5.1. *If $W \in T_\sigma$, then (i) $W' \in \mathcal{X}_\sigma^{\mathsf{snd}}$ iff $W \sqsubseteq W'$; and (ii) $T_\sigma = Q^*[W]$.*

Proof. Suppose $W \in T_\sigma$. For (i), first suppose $W' \in \mathcal{X}_\sigma^{\mathsf{snd}}$ and $W, c \models \mathsf{S}_i\varphi$. Considering the singleton sequence $\delta = \langle i, c, \varphi\rangle$ we have $W \in \mathcal{X}_\delta^{\mathsf{snd}}$. From $W \in T_\sigma$ we get $\delta \preceq \sigma$, i.e. there is $\psi \equiv \varphi$ such that $\langle i, c, \psi\rangle \in \sigma$. From $W' \in \mathcal{X}_\sigma^{\mathsf{snd}}$ and $\mathsf{S}_i\varphi \equiv \mathsf{S}_i\psi$ we get $W', c \models \mathsf{S}_i\varphi$. This shows $W \sqsubseteq W'$.

Now suppose $W \sqsubseteq W'$ and let $\langle i, c, \varphi\rangle \in \sigma$. Then since $W \in T_\sigma \subseteq \mathcal{X}_\sigma^{\mathsf{snd}}$ we have $W, c \models \mathsf{S}_i\varphi$, and $W \sqsubseteq W'$ gives $W', c \models \mathsf{S}_i\varphi$. Consequently $W' \in \mathcal{X}_\sigma^{\mathsf{snd}}$.

Now for (ii), first suppose $W' \in Q^*[W]$. Then W and W' satisfy exactly the same soundness statements, so $W' \in T_\sigma$ also. Conversely, suppose $W' \in T_\sigma$. Then $W' \in \mathcal{X}_\sigma^{\mathsf{snd}}$, so (i) gives $W \sqsubseteq W'$. But we also have $W' \in T_\sigma$ and $W \in \mathcal{X}_\sigma^{\mathsf{snd}}$, so (i) again gives $W' \sqsubseteq W$. Hence $W \approx W'$, i.e. $W' \in Q^*[W]$. □

We can now show that property (6.1) implies **Credulity** together with **Soundness**.

Proposition 6.5.1. *Suppose L satisfies **Soundness** and property (6.1). Then L satisfies **Credulity**.*

Proof. Suppose $T_\sigma, c \not\models \mathsf{S}_i\varphi$. By assumption, there is some $W \in T_\sigma$ such that $W, c \not\models \mathsf{S}_i\varphi$. Take some $W' \in L(\sigma)$. We need to show that $W', c \not\models \mathsf{S}_i\varphi$.

Now, from $W, c \not\models \mathsf{S}_i\varphi$ we get $\Pi_i^W[v_c^W] \cap \|\varphi\| = \emptyset$. Taking ψ such that $\|\psi\| = \Pi_i^W[v_c^W]$, we have $\|\psi\| \cap \|\varphi\| = \emptyset$ and $W, c \models \mathsf{E}_i\psi$. Thus $T_\sigma, c \not\models \neg\mathsf{E}_i\psi$, so property (6.1) gives $L(\sigma), c \models \mathsf{E}_i\psi$. Since $W' \in L(\sigma)$, we get $W', c \models \mathsf{E}_i\psi$, i.e. $\Pi_i^{W'}[\psi] = \|\psi\|$.

On the other hand, from **Soundness** we have $L(\sigma) \subseteq \mathcal{X}_\sigma^{\mathsf{snd}}$, so $W' \in \mathcal{X}_\sigma^{\mathsf{snd}}$. Since $W \in T_\sigma$, Lemma 6.5.1 gives $W \sqsubseteq W'$, and so $\|\psi\| = \Pi_i^W[v_c^W] \subseteq \Pi_i^{W'}[v_c^{W'}]$

[6] We conjecture (6.1) is strictly stronger than **Credulity**.

[7] Indeed, consider the world W such that Π_i^W is the one-cell partition $\{\mathcal{V}\}$ for all sources i. Then $W \in \mathcal{X}_\sigma^{\mathsf{snd}}$ for any σ, since all reports are sound when sources have expertise only on tautologies. Consequently $\mathsf{S}_i\varphi$ is always consistent with $\mathcal{X}_\sigma^{\mathsf{snd}}$, so we can never have $\mathcal{X}_\sigma^{\mathsf{snd}}, c \models \neg\mathsf{S}_i\varphi$. To show **Credulity** fails, it suffices to take any W' such that $W', c \not\models \mathsf{S}_i\varphi$ for some i, c, φ, and to take σ such that $W' \in T_\sigma$.

by Lemma 6.3.1. Now since $\|\psi\|$ is a subset of the cell $\Pi_i^{W'}[v_c^{W'}]$, its expansion under $\Pi_i^{W'}$ is equal to this cell, i.e. $\Pi_i^{W'}[\psi] = \Pi_i^{W'}[v_c^{W'}]$. But we showed above that $\Pi_i^{W'}[\psi] = \|\psi\|$. Hence $\Pi_i^{W'}[v_c^{W'}] = \|\psi\|$. In particular, $\Pi_i^{W'}[v_c^{W'}] \cap \|\varphi\| = \emptyset$, and $W', c \not\models \mathsf{S}_i\varphi$ as required. □

The next two results show that initial segments of streams are (eventually) pseudo-streams, and that any pseudo-stream gives rise to a stream.

Lemma 6.5.2. *If ρ is a stream for W, there is n such that $W \in T_{\rho[m]}$ for all $m \geq n$.*

Proof. Let $\widehat{\cdot}$ be a function which selects a representative formula for each equivalence class of $\mathcal{L}_0/\!\equiv$, so that $\varphi \equiv \widehat{\varphi}$ and $\varphi \equiv \psi$ implies $\widehat{\varphi}$ is equal to $\widehat{\psi}$. Note that since Prop is finite, and since $\mathcal{S}$ and $\mathcal{C}$ are also finite, there are only finitely many reports of the form $\langle i, c, \widehat{\varphi}\rangle$. By completeness of ρ for W, we may take n sufficiently large so that $W, c \models \mathsf{S}_i\widehat{\varphi}$ implies $\langle i, c, \widehat{\varphi}\rangle \in \rho[n]$, for all i, c, φ. Now, take $m \geq n$. We need to show $W \in T_{\rho[m]}$. Clearly $W \in \mathcal{X}^{\mathsf{snd}}_{\rho[m]}$, since ρ is sound for W. Suppose $W \in \mathcal{X}^{\mathsf{snd}}_{\delta}$. We need to show $\delta \preceq \rho[m]$. Indeed, take $\langle i, c, \varphi\rangle \in \delta$. Then $W, c \models \mathsf{S}_i\varphi$. Since $\mathsf{S}_i\varphi \equiv \mathsf{S}_i\widehat{\varphi}$, we have $W, c \models \mathsf{S}_i\widehat{\varphi}$. Hence $\langle i, c, \widehat{\varphi}\rangle$ appears in $\rho[n]$, and consequently in $\rho[m]$ too. Since $\varphi \equiv \widehat{\varphi}$, this shows $\delta \preceq \rho[m]$. □

Lemma 6.5.3. *If $W \in T_\sigma$ and $N = |\sigma|$, there is a stream ρ for W such that $\rho[Nk] \equiv \sigma^k$ for all $k \in \mathbb{N}$.*

Proof. First note that $W \in T_\sigma$ implies $\sigma \neq \emptyset$, so $N > 0$. Since $\mathcal{L}_0$ is countable, we may index the set of $\mathcal{L}_0$ formulas equivalent to $\varphi \in \mathcal{L}_0$ as $\{\varphi_n\}_{n\in\mathbb{N}}$. Let σ_n be obtained from σ by replacing each report $\langle i, c, \varphi\rangle$ with $\langle i, c, \varphi_n\rangle$. Then $\sigma \equiv \sigma_n$. Let ρ be the sequence obtained as the infinite concatenation $\sigma_1 \circ \sigma_2 \circ \sigma_3 \circ \cdots$ (this is possible since σ is of positive finite length). Then $\rho[Nk] = \sigma_1 \circ \cdots \circ \sigma_k$, and consequently $\rho[Nk] \equiv \sigma^k$.

It remains to show ρ is a stream for W. Soundness of ρ follows from $W \in T_\sigma \subseteq \mathcal{X}^{\mathsf{snd}}_\sigma$, since every report in ρ is equivalent to some report in σ by construction. For completeness, suppose $W, c \models \mathsf{S}_i\varphi$. As in the proof of Lemma 6.5.1, considering the singleton sequence $\delta = \langle i, c, \varphi\rangle$, we get from $W \in T_\sigma$ that there is $\psi \equiv \varphi$ such that $\langle i, c, \psi\rangle \in \sigma$. Hence there is $n \in \mathbb{N}$ such that $\varphi = \psi_n$, so $\langle i, c, \varphi\rangle \in \sigma_n$, and thus $\langle i, c, \varphi\rangle \in \rho$. □

Next we obtain an equivalent formulation of **Credulity** which is less transparent as a postulate for learning methods, but easier to work with.

Lemma 6.5.4. *Suppose L satisfies **Soundness**. Then L satisfies **Credulity** if and only if $L(\sigma) \subseteq T_\sigma$ for all σ with $T_\sigma \neq \emptyset$.*

Proof. "if": Suppose $T_\sigma, c \not\models \mathsf{S}_i\varphi$. Then there is $W \in T_\sigma$ such that $W, c \not\models \mathsf{S}_i\varphi$. By our assumption and Lemma 6.5.1, $L(\sigma) \subseteq T_\sigma = Q^*[W]$. Thus every world in $L(\sigma)$ agrees with W on soundness statements, so $L(\sigma), c \models \neg\mathsf{S}_i\varphi$.

"only if": Suppose there is some $W \in T_\sigma$, and take $W' \in L(\sigma)$. We need to show $W' \in T_\sigma$; by Lemma 6.5.1, this is equivalent to $W \approx W'$. First suppose $W, c \models \mathsf{S}_i\varphi$. Then $W \in T_\sigma$ implies there is $\psi \equiv \varphi$ such that $\langle i, c, \psi\rangle \in \sigma$. By **Soundness** for L, we have $W' \in L(\sigma) \subseteq \mathcal{X}^{\mathsf{snd}}_\sigma$. Consequently $W', c \models \mathsf{S}_i\psi$ and thus $W', c \models \mathsf{S}_i\varphi$. This shows $W \sqsubseteq W'$. Now suppose $W, c \not\models \mathsf{S}_i\varphi$. Then $T_\sigma, c \not\models \mathsf{S}_i\varphi$. By **Credulity**, $L(\sigma), c \models \neg\mathsf{S}_i\varphi$. Hence $W', c \not\models \mathsf{S}_i\varphi$. This shows $W' \sqsubseteq W$. Thus $W \approx W'$ as required. □

Finally, we prove the characterisation of truth-tracking.

Proof of Theorem 6.5.1. Suppose L satisfies **Equivalence**, **Repetition** and **Soundness**.

"if": Suppose **Credulity** holds. We show L solves Q^*. Take any world W and stream ρ for W. By Lemma 6.5.2, there is n such that $W \in T_{\rho[m]}$ for all $m \geq n$. By Lemma 6.5.1, $T_{\rho[m]} = Q^*[W]$ for such m. In particular, $T_{\rho[m]} \neq \emptyset$. By **Credulity** and Lemma 6.5.4, we get $L(\rho[m]) \subseteq T_{\rho[m]} = Q^*[W]$.

"only if": Suppose L solves Q^*. We show **Credulity** via Lemma 6.5.4. Suppose there is some $W \in T_\sigma$, and write $N = |\sigma| > 0$. By Lemma 6.5.3, there is a stream ρ for W such that $\rho[Nk] \equiv \sigma^k$ for all $k \in \mathbb{N}$. By **Repetition** and **Equivalence**, $L(\sigma) = L(\sigma^k) = L(\rho[Nk])$. But L solves Q^*, so for k sufficiently large we have $L(\rho[Nk]) \subseteq Q^*[W] = T_\sigma$. Hence, going via some large k, we obtain $L(\sigma) \subseteq T_\sigma$ as required. □

6.5.2 Conditioning Methods

In this section we turn to the family of *conditioning* methods from Section 5.3.1. While the interpretation of input sequences is different when considering sound and complete streams – we read $\langle i, c, \varphi\rangle$ as i reporting φ is *possible* in case c, whereas no such fixed interpretation was in place before – this class of methods can still be applied. Moreover, while conditioning methods were put forward in Section 5.3.1 in order to satisfy rationality postulates, we will soon see that they are also compatible with truth-tracking.

Conditioning methods operate by successively restricting a fixed *plausibility total preorder* to the information corresponding to each new report $\langle i, c, \varphi\rangle$. In this section, we take a report $\langle i, c, \varphi\rangle$ to correspond to the information that $\mathsf{S}_i\varphi$ holds in case c; this fits with our assumption throughout that sources only report sound statements.[8] Thus, the worlds under consideration given a sequence σ are exactly those satisfying all soundness statements in σ, i.e. $\mathcal{X}^{\mathsf{snd}}_\sigma$. Note that $\mathcal{X}^{\mathsf{snd}}_\sigma$ represents the *indefeasible knowledge* given by σ: worlds outside $\mathcal{X}^{\mathsf{snd}}_\sigma$ are eliminated and cannot be recovered with further reports, since $\mathcal{X}^{\mathsf{snd}}_{\sigma\circ\delta} \subseteq \mathcal{X}^{\mathsf{snd}}_\sigma$. The plausibility order allows us to represent *defeasible beliefs* about the most plausible worlds within $\mathcal{X}^{\mathsf{snd}}_\sigma$.

[8]This is more restrictive than in the previous chapter, where the knowledge component of an operator enjoyed more freedom. Our assumption here is equivalent to fixing $K^\sigma = \mathrm{Cn}(G^\sigma_{\mathsf{snd}})$.

Definition 6.5.1. *For a total preorder* $\leq$ *on* $\mathcal{W}$*, the* conditioning method $L_{\leq}$ *is given by* $L_{\leq}(\sigma) = \min_{\leq} \mathcal{X}_{\sigma}^{\mathsf{snd}}$*.*

Note that since $\mathcal{X}_{\sigma}^{\mathsf{snd}} \neq \emptyset$ for all σ[9] and $\mathcal{W}$ is finite, $L_{\leq}$ is consistent. Moreover, $L_{\leq}$ satisfies **Equivalence**, **Repetition** and **Soundness** for any choice of $\leq$.

Example 6.5.1. *We recall two concrete choices of* $\leq$ *from Chapter 5.*

1. *Set* $W \leq W'$ *iff* $r(W) \leq r(W')$*, where*

$$r(W) = -\sum_{i \in \mathcal{S}} |\{p \in \mathsf{Prop} \mid \Pi_i^W[p] = \|p\|\}|.$$

The most plausible worlds in this order are those in which source have as much expertise on the propositional variables as possible, on aggregate. The corresponding conditioning method is denoted by L_{vbc}*, standing for* variable-based conditioning.

2. *Set* $W \leq W'$ *iff* $r(W) \leq r(W')$*, where*

$$r(W) = -\sum_{i \in \mathcal{S}} |\Pi_i^W|.$$

This order aims to maximise the number of cells in each source's partitions, thereby maximising the number of propositions on which they have expertise. Note that the propositional variables play no special role. The corresponding conditioning operator is denoted by L_{pbc}*, for* partition-based conditioning.

A straightforward property of $\leq$ characterises truth-tracking for conditioning methods. For a generic total preorder $\leq$, let $<$ denote its strict part.

Theorem 6.5.2. $L_{\leq}$ *is truth-tracking if and only if*

$$W \sqsubset W' \implies \exists W'' \approx W \text{ such that } W'' < W'. \tag{6.2}$$

Like **Credulity**, (6.2) is a principle of maximising trust in sources. Recall from that Lemma 6.3.1 that $W \sqsubset W'$ means all sources are more knowledgeable in each case in W than in W', and there is at least one source and case for which this holds strictly. If we aim to trust sources as much as possible, we might impose $W < W'$ here; then W' is strictly less plausible and will be ruled out in favour of W. This yields a sufficient condition for truth-tracking, but to obtain a necessary condition we need to allow a "surrogate" world $W'' \approx W$ to take the place of W.

(6.2) is also intuitively similar to **Refinement** from Chapter 5, which requires that $W \leq W'$ whenever $W \preceq W'$ (where the latter condition means each partition Π_i^W refines $\Pi_i^{W'}$), but with the strict part of $\sqsubseteq$ taking the place of $\preceq$.

[9] For example, if $\Pi_i^W = \{\mathcal{V}\}$ for all i then $W \in \mathcal{X}_{\sigma}^{\mathsf{snd}}$ for all σ.

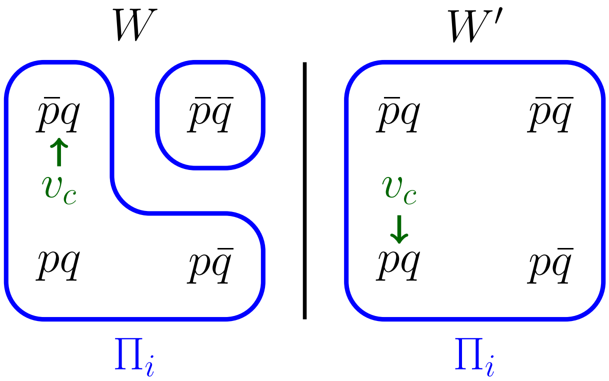


Figure 6.3: Worlds which demonstrate $\mathsf{L}_{\mathsf{vbc}}$ *is not truth-tracking.*

Proof of Theorem 6.5.2. Write $L = L_{\leq}$. Since L satisfies **Equivalence**, **Repetition** and **Soundness**, we may use Theorem 6.5.1. Furthermore, it is sufficient by Lemma 6.5.4 to show that (6.2) holds if and only if $L(\sigma) \subseteq T_\sigma$, whenever $T_\sigma \neq \emptyset$.

"if": Suppose $W \sqsubset W'$. Let σ be some pseudo-stream for W, so that $W \in T_\sigma$.[10] Note that since $W \in T_\sigma \subseteq \mathcal{X}_\sigma^{\mathsf{snd}}$ and $W \sqsubset W'$, we have $W' \in \mathcal{X}_\sigma^{\mathsf{snd}}$ also. By assumption, $L(\sigma) \subseteq T_\sigma = Q^*[W]$. Since $W \not\approx W'$, this means $W' \in \mathcal{X}_\sigma^{\mathsf{snd}} \setminus L(\sigma)$. That is, W' lies in $\mathcal{X}_\sigma^{\mathsf{snd}}$ but is not $\leq$-minimal. Consequently there is $W'' \in \mathcal{X}_\sigma^{\mathsf{snd}}$ such that $W'' < W'$. Since L is consistent, we may assume without loss of generality that $W'' \in L(\sigma)$. Hence $W'' \in Q^*[W]$, so $W'' \approx W$.

"only if": Suppose there is some $W \in T_\sigma$, and let $W' \in L(\sigma)$. We need to show $W' \in T_\sigma = Q^*[W]$, i.e. $W \approx W'$. Since $W' \in L(\sigma) \subseteq \mathcal{X}_\sigma^{\mathsf{snd}}$, Lemma 6.5.1 gives $W \sqsubseteq W'$. Suppose for contradiction that $W \not\approx W'$. Then $W \sqsubset W'$. By (6.2), there is $W'' \approx W$ such that $W'' < W'$. But W' is $\leq$-minimal in $\mathcal{X}_\sigma^{\mathsf{snd}}$, so this must mean $W'' \notin \mathcal{X}_\sigma^{\mathsf{snd}}$. On the other hand, $W'' \in Q^*[W] = T_\sigma \subseteq \mathcal{X}_\sigma^{\mathsf{snd}}$: contradiction. □

Example 6.5.2. *We revisit the methods of Example 6.5.1.*

1. *The variable-based conditioning method* $\mathsf{L}_{\mathsf{vbc}}$ *is* not *truth-tracking. Indeed, consider the worlds W and W' shown in Fig. 6.3, where we assume* $\mathsf{Prop} = \{p, q\}$, $\mathcal{S} = \{i\}$ *and* $\mathcal{C} = \{c\}$. *Then $W \sqsubset W'$ (e.g. by Lemma 6.3.1). Note that i does not have expertise on p or q in both W and W', so $r(W) = r(W') = 0$. Moreover, i's partition is uniquely determined in $Q^*[W]$ by Theorem 6.4.3, so if $W'' \approx W$ then $r(W'') = 0$ also. That is, there is no $W'' \approx W$ such that $W'' < W'$. Hence (6.2) fails, and* $\mathsf{L}_{\mathsf{vbc}}$ *is not truth-tracking. Intuitively, the problem here is that since i's expertise is not split along the lines of the propositional variables when W is the actual world,* $\mathsf{L}_{\mathsf{vbc}}$ *will always maintain W' as a possibility.*

[10]For example, pick some stream ρ and apply Lemma 6.5.2 to obtain a pseudo-stream.

2. *The partition-based conditioning method* L_{pbc} is *truth-tracking. Indeed, if* $W \sqsubset W'$ *we may construct* W'' *from* W *by modifying the partition of each source* i *so that all valuations outside of* $\bigcup_{c \in \mathcal{C}} \Pi_i^W[v_c^W]$ *lie in their own cell. Then* $W \approx W''$*. One can show that* $\Pi_i^{W''}$ *refines* $\Pi_i^{W'}$ *for all* $i \in \mathcal{S}$*, and there is some* i *for which the refinement is strict. Hence the partitions in* W'' *contain strictly more cells, so* $W'' < W'$*.*

6.5.3 Score-based Methods

In this section we consider the other class of methods introduced in Chapter 5: *score-based* methods. As with conditioning methods above, we consider a more restricted class in which we dispense with the prior plausibility ranking r_0 from Definition 5.3.4, and fix knowledge as soundness.

Definition 6.5.2. *For a function* $d : \mathcal{W} \times (\mathcal{S} \times \mathcal{C} \times \mathcal{L}_0) \to \mathbb{N}_0$ *and a sequence* σ*, write*

$$r_d^\sigma(W) = \sum_{\langle i,c,\varphi \rangle \in \sigma} d(W, \langle i, c, \varphi \rangle).$$

The score-based method L_d *is then given by* $L_d(\sigma) = \mathrm{argmin}_{W \in \mathcal{X}_\sigma^{\mathsf{snd}}} r_d^\sigma(W)$*.*

As before, $d(W, \langle i, c, \varphi \rangle)$ represents a measure of "disagreement" between the world W and report $\langle i, c, \varphi \rangle$; the greater $d(W, \langle i, c, \varphi \rangle)$, the less plausible it is deemed for i to report φ in case c. The conjecture $L_d(\sigma)$ consists of the worlds W satisfying the soundness constraints of σ with minimal disagreement score, computed as the sum $r_d^\sigma(W)$ of the disagreement on each report. L_d is consistent and satisfies both **Repetition** and **Soundness** for any choice of d.

Example 6.5.3. *Adapting the score-based example from Chapter 5 to this setting, take*

$$d(W, \langle i, c, \varphi \rangle) = |\Pi_i^W[\varphi] \setminus \|\varphi\||.$$

The corresponding method aims to minimise the "excess" valuations in $\Pi_i^W[\varphi]$ *which are not themselves models of* φ*. We denote it by* L_{exm}*, standing for* excess-minimisation*.*

Truth-tracking for score-based operators satisfying **Equivalence** can be characterised in almost exactly the same way as for conditioning operators, using a property similar to (6.2).

Theorem 6.5.3. *Suppose* d *is such that* $d(W, \langle i, c, \varphi \rangle) = d(W, \langle i, c, \psi \rangle)$ *whenever* $\varphi \equiv \psi$*. Then* L_d *is truth-tracking if and only if*

$$W \in T_\sigma \text{ and } W \sqsubset W' \implies \exists W'' \approx W \text{ such that } r_d^\sigma(W'') < r_d^\sigma(W'). \tag{6.3}$$

The proof is essentially identical to that of Theorem 6.5.2, and is thus omitted.

Example 6.5.4. *Revisiting Example 6.5.3, we find that* L_{exm} *is truth-tracking. Indeed, it is clear that* d *treats equivalent formula identically, since* $d(W, \langle i, c, \varphi\rangle)$ *only depends on* Π_i^W *and* $\|\varphi\|$. *Given* $W \in T_\sigma$ *and* $W \sqsubset W'$ *one can take* W'' *in the same way as for* L_{pbc} *in Example 6.5.2. Then* $\Pi_i^{W''}$ *refines* $\Pi_i^{W'}$ *for all* i, *so* $d(W'', \langle i, c, \varphi\rangle) \leq d(W', \langle i, c, \varphi\rangle)$ *for all* $\langle i, c, \varphi\rangle \in \sigma$. *Consequently,* $r_d^\sigma(W'') \leq r_d^\sigma(W')$. *Moreover, since* $W \sqsubset W'$ *there is some* i *and* c *such that* $\Pi_i^W[v_c^W] \subset \Pi_i^{W'}[v_c^{W'}]$. *Taking any* φ *such that* $\|\varphi\| = \Pi_i^W[v_c^W]$, *we have* $W, c \models \mathsf{S}_i\varphi$. *Since* $W \in T_\sigma$, *there is some* $\psi \equiv \varphi$ *such that* $\langle i, c, \psi\rangle \in \sigma$. *Now, since* $W \approx W''$ *we have* $\Pi_i^{W''}[v_c^{W''}] = \Pi_i^W[v_c^W]$. *Consequently*

$$\Pi_i^{W''}[\psi] = \Pi_i^{W''}[\Pi_i^W[v_c^W]] = \Pi_i^{W''}[\Pi_i^{W''}[v_c^{W''}]] = \Pi_i^{W''}[v_c^{W''}] = \Pi_i^W[v_c^W] = \|\psi\|$$

and thus $d(W'', \langle i, c, \psi\rangle) = 0$. *On the other hand,* $\Pi_i^{W'}[v_c^{W'}] \supset \|\psi\|$ *implies* $\Pi_i^{W'}[\psi] = \Pi_i^{W'}[v_c^{W'}]$, *and so*

$$d(W', \langle i, c, \psi\rangle) = |\Pi_i^{W'}[v_c^{W'}] \setminus \|\psi\|| > 0 = d(W'', \langle i, c, \psi\rangle),$$

which gives $r_d^\sigma(W'') < r_d^\sigma(W')$ *as required.*

6.6 Belief-based streams

So far our model of source inputs is based on sound and complete streams, in which sources report all (and only) sound formulas. This is perhaps not realistic in situations where sources have *beliefs*, possibly going beyond their knowledge. For instance, in our running example the doctor D may *believe* A suffers from p but not *know* so; in this case $\neg p$ is sound – and would thus be reported, in our model so far – when in fact D believes its negation. To remedy this we introduce additional structure on top of worlds which express sources' beliefs about the true valuation in each case. The analogue of a stream then enumerates the beliefs of each source. Such streams differ from the earlier notion in that while all beliefs are sound, not all sound statements are believed. Intuitively such streams are less informative; this will be formalised in Proposition 6.6.2 below.

Formally, each source's beliefs in a particular case will be modelled by *plausibility orders* over valuations.

Definition 6.6.1. *A* belief structure *is an assignment* $\mathcal{B} = \{\leq_{ic}\}_{i \in \mathcal{S}, c \in \mathcal{C}}$ *of total preorders* $\leq_{ic}$ *over* $\mathcal{V}$ *to each source* i *and case* c.

A belief structure is independent of any world, and can be thought of as the source's *prior beliefs* about the "actual" valuation in each case: $u \leq_{ic} v$ means i considers u at least as plausible as v for the actual valuation in case c. In a given world W, the set $\Pi_i^W[v_c^W]$ contains the valuations indistinguishable from the actual valuation v_c^W. This can be interpreted as the *observation* of i for case c in the world W, since it contains exactly those valuations which are possible candidates from the point of view of i's expertise. Combining a belief structure $\mathcal{B}$ with a world W, the maximally plausible such candidates form each source's belief set.

Definition 6.6.2. *Given a belief structure $\mathcal{B} = \{\leq_{ic}\}_{i\in\mathcal{S},c\in\mathcal{C}}$ and a world W, source i* believes *φ in case c, denoted $W, c \models_{\mathcal{B}} \mathsf{B}_i\varphi$, if*

$$\min_{\leq_{ic}} \Pi_i^W[v_c^W] \subseteq \|\varphi\|.$$

That is, i believes φ in case c if φ holds in all the most plausible states among those indistinguishable from the actual one. Since $\mathcal{V}$ is finite and $\Pi_i^W[v_c^W]$ is non-empty (it contains v_c^W, at least), so too is the minimum. Consequently, the beliefs of each source in each case are consistent. There is also similarity with the conditioning methods from Section 6.5.2, where a plausibility order is used to select the most plausible worlds. A key difference here is that each source has a plausibility order over *valuations*, not worlds.

This notion of belief is compatible with expertise and soundness: all beliefs are sound, and experts hold correct beliefs.

Proposition 6.6.1. *Let $\mathcal{B}$ be a belief structure and W a world. Then*

1. *If $W, c \models_{\mathcal{B}} \mathsf{B}_i\varphi$ then $W, c \models \mathsf{S}_i\varphi$.*
2. *If $W, c \models \mathsf{E}_i\varphi$, then $W, c \models_{\mathcal{B}} \mathsf{B}_i\varphi$ if and only if $W, c \models \varphi$.*

Proof. Fix $\mathcal{B}$ and W.

1. Suppose $W, c \models_{\mathcal{B}} \mathsf{B}_i\varphi$. As noted above, $\min_{\leq_{ic}} \Pi_i^W[v_c^W]$ is non-empty, and thus contains some valuation v. Since $\mathsf{B}_i\varphi$ holds, we have $v \in \|\varphi\|$. But also $v \in \Pi_i^W[v_c^W]$; hence $v_c^W \in \Pi_i^W[v] \subseteq \Pi_i^W[\varphi]$, i.e. $W, c \models \mathsf{S}_i\varphi$.

2. Suppose $W, c \models \mathsf{E}_i\varphi$. First suppose $W, c \models_{\mathcal{B}} \mathsf{B}_i\varphi$. By (1), $W, c \models \mathsf{S}_i\varphi$. Since $\mathsf{E}_i\varphi \wedge \mathsf{S}_i\varphi \to \varphi$ is a validity, $W, c \models \varphi$ as required. Conversely, suppose $W, c \models \varphi$. Then $v_c^W \in \|\varphi\|$, so $\Pi_i^W[v_c^W] \subseteq \Pi_i^W[\varphi]$. But since i has expertise on φ, we have $\Pi_i^W[\varphi] = \|\varphi\|$; hence $\Pi_i^W[v_c^W] \subseteq \|\varphi\|$. Clearly the $\leq_{ic}$-minimal valuations in $\Pi_i^W[v_c^W]$ are also members of $\|\varphi\|$, so $W, c \models_{\mathcal{B}} \mathsf{B}_i\varphi$.

□

We come to an example.

Example 6.6.1. *Consider the setting of Example 6.0.1. Based on the age and other characteristics of the patients A and B, the doctor D has prior beliefs about the likely incidence of conditions p and q, as described by the following belief structure $\mathcal{B}$:*

$$\leq_{\mathsf{D},A} : \bar{p}\bar{q} < p\bar{q} < \bar{p}q < pq$$
$$\leq_{\mathsf{D},B} : pq \simeq \bar{p}q \simeq p\bar{q} \simeq \bar{p}\bar{q}.$$

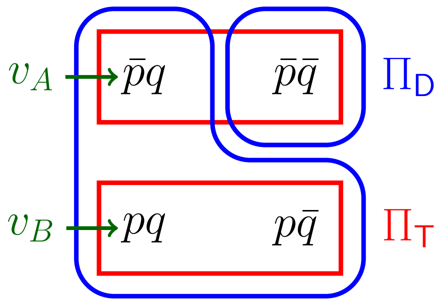


Figure 6.4: Reproduction of Fig. 6.1.

That is, D *considers it on the whole plausible for* A *to be healthy on both counts, but more likely that* A *suffers only from* p *than only* q*. On the other hand, no data is available for patients of* B*'s age, and thus the plausibility order* $\leq_{\mathsf{D},B}$ *is the flat one in which all possibilities are equally plausible.*

Supposing W *from Fig. 6.1 is the actual world (for convenience we reproduce the figure in Fig. 6.4),* D *then observes both patients. Since the actual valuations* v_A^W *and* v_B^W *lie in the same cell in* Π_{D}^W*,* D *lacks expertise to distinguish them with certainty; we have*

$$\Pi_{\mathsf{D}}^W[v_A^W] = \Pi_{\mathsf{D}}^W[v_B^W] = \{pq, p\bar{q}, \bar{p}q\}.$$

However, the difference in the plausibility orders $\leq_{\mathsf{D},A}$ *and* $\leq_{\mathsf{D},B}$ *leads to different beliefs: we have*

$$\min\nolimits_{\leq_{\mathsf{D},A}} \Pi_{\mathsf{D}}^W[v_A^W] = \min\nolimits_{\leq_{\mathsf{D},A}}\{pq, p\bar{q}, \bar{p}q\} = \{p\bar{q}\}$$

whereas

$$\min\nolimits_{\leq_{\mathsf{D},B}} \Pi_{\mathsf{D}}^W[v_B^W] = \Pi_{\mathsf{D}}^W[v_B^W] = \{pq, p\bar{q}, \bar{p}q\}.$$

Note that for A *the doctor* D *has beliefs going beyond mere knowledge, and in fact* D *holds* false beliefs*: we have* $W, A \models_{\mathcal{B}} \mathsf{B}_{\mathsf{D}}p$ *but* $W, A \models \neg p$*. This is not so for* B*; in this case all beliefs are known and therefore true, since the actual valuation* $v_B^W = pq$ *is contained in* $\min_{\leq_{\mathsf{D},B}} \Pi_{\mathsf{D}}^W[v_B^W]$*.*

We can now consider "belief-based" streams, which enumerate source beliefs.

Definition 6.6.3. *A sequence of reports* ρ *is a* $\mathcal{B}$-stream *for a world* W *iff for all* i, c, φ*:*

$$\langle i, c, \varphi\rangle \in \rho \iff W, c \models_{\mathcal{B}} \mathsf{B}_i\varphi.$$

This is essentially the same notion as streams from Definition 6.2.1, but with $\mathsf{B}_i\varphi$ replacing $\mathsf{S}_i\varphi$.

Example 6.6.2. *Consider* W *from Fig. 6.4 and source* D*. Continuing from Example 6.6.1, the reports from* D *for case* A *in any* $\mathcal{B}$*-stream enumerate* $\mathrm{Cn}_0(p \wedge \neg q)$*, and the reports for case* B *enumerate* $\mathrm{Cn}_0(p \vee q)$*.*

A notion of $\mathcal{B}$-*solvability* comes naturally.

Definition 6.6.4. *A method L is said to $\mathcal{B}$-*solve *a question Q if for all W and all $\mathcal{B}$-streams ρ for W, there is $n \in \mathbb{N}$ such that $L(\rho[m]) \subseteq Q[W]$ for all $m \geq n$.*

Again mirroring the definitions from earlier, a preorder $\sqsubseteq_{\mathcal{B}}$ can be defined on worlds, where

$$W \sqsubseteq_{\mathcal{B}} W' \iff \forall i, c, \varphi : \ (W, c \models_{\mathcal{B}} \mathsf{B}_i\varphi \implies W', c \models \mathsf{B}_i\varphi).$$

The analogue of Lemma 6.3.1 is as follows.

Lemma 6.6.1. *$W \sqsubseteq_{\mathcal{B}} W'$ if and only if for all $i \in \mathcal{S}$ and $c \in \mathcal{C}$, $\min_{\leq_{ic}} \Pi_i^W[v_c^W] \supseteq \min_{\leq_{ic}} \Pi_i^{W'}[v_c^{W'}]$.*

Proof. The "if" direction is straightforward. For the "only if" part, suppose $W \sqsubseteq_{\mathcal{B}} W'$, and take $i \in \mathcal{S}$, $c \in \mathcal{C}$. Taking φ to be any formula with $\|\varphi\| = \min_{\leq_{ic}} \Pi_i^W[v_c^W]$, we have $W, c \models_{\mathcal{B}} \mathsf{B}_i\varphi$. Since $W \sqsubseteq_{\mathcal{B}} W'$ this means $W', c \models_{\mathcal{B}} \mathsf{B}_i\varphi$, i.e.

$$\min_{\leq_{ic}} \Pi_i^{W'}[v_c^{W'}] \subseteq \|\varphi\| = \min_{\leq_{ic}} \Pi_i^W[v_c^W]$$

as required. □

The equivalence relation $\approx_{\mathcal{B}}$ induced by $\sqsubseteq_{\mathcal{B}}$ defines a question $Q^*_{\mathcal{B}}$. As one might anticipate by our choice of notation, $Q^*_{\mathcal{B}}$ is the hardest $\mathcal{B}$-solvable question. In analogy with the $\mathcal{X}_\sigma^{\mathsf{snd}}$ notation for the set of worlds satisfying all soundness constraints of σ, write $\mathcal{X}_\sigma^{\mathcal{B}}$ for the set of worlds W such that $W, c \models_{\mathcal{B}} \mathsf{B}_i\varphi$ for all $\langle i, c, \varphi \rangle \in \sigma$.

Theorem 6.6.1. *Q is $\mathcal{B}$-solvable if and only if $Q^*_{\mathcal{B}} \preceq Q$.*

Proof (sketch). "if": It suffices to show that $Q^*_{\mathcal{B}}$ can be $\mathcal{B}$-solved by a consistent method L. In fact, the same approach to showing Q^* is solvable on sound streams (Proposition 6.3.3) can be applied here. Set $L(\sigma) = \min_{\sqsubseteq_{\mathcal{B}}} \mathcal{X}_\sigma^{\mathcal{B}}$ if $\mathcal{X}_\sigma^{\mathcal{B}} \neq \emptyset$, and $L(\sigma) = \mathcal{W}$ otherwise. By an argument identical to that of Proposition 6.3.3 – with notions of soundness replaced by belief with respect to $\mathcal{B}$ – one can show L solves $Q^*_{\mathcal{B}}$.

"only if": This can be shown by an argument identical to that of Lemma 6.3.2, using the fact that $W \approx_{\mathcal{B}} W'$ implies W and W' have exactly the same $\mathcal{B}$-streams. □

The question of whether there is a *single* method L which simulatenously $\mathcal{B}$-solves all $Q^*_{\mathcal{B}}$ remains open.

Theorem 6.6.1 shows that $Q^*_{\mathcal{B}}$ plays the same role with respect to $\mathcal{B}$-streams as Q^* does with respect to sound and complete streams. Accordingly, we say a method L is $\mathcal{B}$-*truth-tracking* if it $\mathcal{B}$-solves $Q^*_{\mathcal{B}}$. By comparing Q^* with $Q^*_{\mathcal{B}}$, one can assess the *informativeness* of truth-tracking via sound and complete streams versus $\mathcal{B}$-streams. We shall see that Q^* is always at least as informative as $Q^*_{\mathcal{B}}$: we have $Q^* \preceq Q^*_{\mathcal{B}}$. This can be interpreted as informativeness via the analysis of Section 6.4, where we looked at what one can learn about W from the correct answer $Q^*[W]$ (which truth-tracking methods are guaranteed to eventually find). Indeed, $Q^* \preceq Q^*_{\mathcal{B}}$ means $Q^*[W] \subseteq Q^*_{\mathcal{B}}[W]$ for all W, so the correct answer to Q^* narrows down the possibilities for the "actual" world to a greater extent than $Q^*_{\mathcal{B}}$ does, and thus reveals more information about the actual world itself.

Moreover, Q^* is *strictly* more informative for all but one choice of belief structure: the "flat" structure $\mathcal{B}_\flat$, where each $\leq_{ic}$ is the flat ranking $\leq_{ic} = \mathcal{V} \times \mathcal{V}$. In this case source beliefs do not go beyond knowledge at all; all reports in a $\mathcal{B}_\flat$-stream are therefore *true*, but will not be very *specific* when sources are non-experts.

Proposition 6.6.2. *$Q^* \preceq Q^*_{\mathcal{B}}$ for all belief structures $\mathcal{B}$, and $Q^*_{\mathcal{B}} = Q^*$ if and only if $\mathcal{B} = \mathcal{B}_\flat$.*[11]

Proof. First, take any belief structure $\mathcal{B}$. We show that $Q^*[W] \subseteq Q^*_{\mathcal{B}}[W]$ for all W. Indeed, if $W' \in Q^*[W]$ then $W \approx W'$, so by Lemma 6.3.1 we have $\Pi_i^W[v_c^W] = \Pi_i^{W'}[v_c^{W'}]$ for all i and c. Evidently the minimal elements with respect to $\leq_{ic}$ of both sets must also be the same, so we get $W \approx_{\mathcal{B}} W'$ by Lemma 6.6.1. That is, $W' \in Q^*_{\mathcal{B}}[W]$. Hence $Q^* \preceq Q^*_{\mathcal{B}}$.

For the second statement, we first show the "if" direction, i.e. that $Q^*_{\mathcal{B}_\flat} = Q^*$. Indeed, since $\leq_{ic} = \mathcal{V} \times \mathcal{V}$ we have $\min_{\leq_{ic}} U = U$ for any set $U \subseteq \mathcal{V}$. Consequently $W \approx_{\mathcal{B}_\flat} W'$ if and only if $\Pi_i^W[v_c^W] = \Pi_i^{W'}[v_c^{W'}]$ – using Lemma 6.6.1 – which is in turn equivalent to $W \approx W'$ by Lemma 6.3.1. Hence $Q^*_{\mathcal{B}_\flat} = Q^*$.

For the "only if" part we show the contrapositive. Suppose $\mathcal{B} \neq \mathcal{B}_\flat$. We will show $Q^*_{\mathcal{B}} \not\preceq Q^*$, i.e. that there are W, W' such that $W \approx_{\mathcal{B}} W'$ but $W \not\approx W'$. Since $\mathcal{B} \neq \mathcal{B}_\flat$, there exist $i_0 \in \mathcal{S}$, $c_0 \in \mathcal{C}$ and $u, v \in \mathcal{V}$ such that $u <_{i_0,c_0} v$. By the assumption $|\mathsf{Prop}| \geq 2$ we have $|\mathcal{V}| \geq 4$, so we may choose some $w \in \mathcal{V} \setminus \{u, v\}$. Construct W by setting

$$\begin{aligned} v_{c_0}^W &= u, \\ v_c^W &= w, \qquad (c \in \mathcal{C} \setminus \{c_0\}) \\ \Pi_i^W &= \Pi_\perp. \qquad (i \in \mathcal{S}) \end{aligned}$$

Then $\Pi_i^W[v_c^W]$ is the singleton set $\{v_c^W\}$ for all i and c, and so too is belief set $\min_{\leq_{ic}} \Pi_i^W[v_c^W]$. Now construct W' by modifying the partition of source i_0 so that

$$\Pi_{i_0}^{W'} = \{\{u, v\}, \{w\}, \mathcal{V} \setminus \{u, v, w\}\}.$$

The belief sets of sources $i \neq i_0$ are clearly unchanged, as are the belief sets for i_0 in cases $c \neq c_0$ (since $\Pi_{i_0}^{W'}[v_c^{W'}]$ is the singleton set $\{w\}$). For i_0 in case c_0 we have $\Pi_{i_0}^{W'}[v_{c_0}^{W'}] = \Pi_{i_0}^{W'}[u] = \{u, v\}$; but since $u <_{i_0,c_0} v$ by assumption,

$$\min_{\leq_{i_0,c_0}} \Pi_{i_0}^{W'}[v_{c_0}^{W'}] = \{u\} = \min_{\leq_{i_0,c_0}} \Pi_{i_0}^{W}[v_{c_0}^{W}].$$

[11] For the second statement we assume $|\mathsf{Prop}| \geq 2$.

From Lemma 6.6.1 it follows that $W \approx_{\mathcal{B}} W'$. However, we have already seen that $\Pi_{i_0}^W[v_{c_0}^W] = \{u\}$ whereas $\Pi_{i_0}^{W'}[v_{c_0}^{W'}] = \{u, v\}$, so $W \not\approx W'$ by Lemma 6.3.1. This completes the proof. □

Both parts of Proposition 6.6.2 together imply $Q^*_{\mathcal{B}_\flat} \preceq Q^*_{\mathcal{B}}$. However, some subtelty is required in interpreting this refinement. In Section 6.3 refinement of questions was expressed as a *difficulty ordering*, since if $Q \preceq Q'$ then any method solving Q will also solve Q'; in this sense Q' is *easier* than Q. This also holds in the extension to belief-based streams, but only for a *fixed* notion of solvability. For instance, $Q^*_{\mathcal{B}_\flat} \preceq Q^*_{\mathcal{B}}$ does *not* imply that a method $\mathcal{B}_\flat$-solving $Q^*_{\mathcal{B}_\flat}$ will $\mathcal{B}$-solve $Q^*_{\mathcal{B}}$, i.e. that $\mathcal{B}_\flat$-truth-tracking methods are $\mathcal{B}$-truth-tracking for all other $\mathcal{B}$.[12]

Rather, Proposition 6.6.2 places an upper bound on the informativeness of truth-tracking via belief-based streams. Indeed, in certain instances belief-based streams are particularly *uninformative*. For example, consider W such that all sources have complete expertise, i.e. $\Pi_i^W = \Pi_\perp$ for all $i \in \mathcal{S}$. Under soundness-based streams, the actual valuation v_c^W is determined exactly in $Q^*[W]$, and is thus found by truth-tracking methods. However, one can easily construct a belief structure $\mathcal{B}$ and an alternative world W' such that all sources share the same beliefs in W as in W' – i.e. $W \approx_{\mathcal{B}} W'$ – but where these beliefs are *false* in W'. Consequently, $Q^*_{\mathcal{B}}[W]$ reveals no information about the actual valuations whatsoever, despite the fact that the sources are unanimously correct in W.

The problem here is the existence of "degenerate" worlds such as W', where all sources are unanimously incorrect in their beliefs. In probability-based Jury theorems (Grofman, Owen, and Feld 1983), such degenerate worlds occur with low probability due to the choice of the probabilistic model and assumption placed on sources. We sketch a somewhat more extreme solution, where such worlds are excluded from consideration entirely.

Concretely, given a belief structure and a set of worlds $S \subseteq \mathcal{W}$ and $W \in S$, thought of as the "actual" world, one can consider the extent to which $Q^*_{\mathcal{B}}[W] \cap S$ reveals information about W. That is, worlds in $\mathcal{W} \setminus S$ are ignored. In place of a systemic study, we explore this concept via examples of particular choices of S. In what follows we consider an arbitrary but fixed belief structure $\mathcal{B}$.

A majority of experts. For a given formula $\varphi \in \mathcal{L}_0$, let S_φ denote the set of worlds in which a strict majority of sources have expertise on φ:

$$W \in S_\varphi \iff |\{i \in \mathcal{S} \mid \Pi_i^W[\varphi] = \|\varphi\|\}| > \frac{|\mathcal{S}|}{2}.$$

[12] Indeed, this is false: since all reports in a $\mathcal{B}_\flat$-stream are true, they must be jointly consistent in each case $c \in \mathcal{C}$. One can therefore construct a $\mathcal{B}_\flat$-truth-tracking method L which reverts to the trivial conjecture $L(\sigma) = \mathcal{W}$ whenever the reports in σ are not jointly consistent; this problem never arises on $\mathcal{B}_\flat$-streams, but will cause L to fail to be $\mathcal{B}$-truth-tracking for non-flat $\mathcal{B}$.

Then for $W \in S_\varphi$, $Q^*_{\mathcal{B}}[W] \cap S_\varphi$ decides φ in all cases $c \in \mathcal{C}$.

Indeed, take $c \in \mathcal{C}$. Letting I_W denote the set of sources with expertise on φ in W, for any two worlds $W, W' \in S_\varphi$ there is necessarily some $i \in I_W \cap I_{W'}$; otherwise $|I_W \cup I_{W'}| = |I_W| + |I_{W'}| > |\mathcal{S}|$, which cannot be. Thus, $W, c \models \mathsf{E}_i\varphi$ and $W', c \models \mathsf{E}_i\varphi$. If additionally $W \approx_{\mathcal{B}} W'$ then $W, c \models_{\mathcal{B}} \mathsf{B}_i\varphi$ if and only if $W', c \models_{\mathcal{B}} \mathsf{B}_i\varphi$. But by Proposition 6.6.1 (2), this means $W, c \models \varphi$ if and only if $W', c \models \varphi$. That is to say, each world in $Q^*_{\mathcal{B}}[W] \cap S_\varphi$ agrees with W on the truth value of φ in case c, and consequently $Q^*_{\mathcal{B}}[W] \cap S_\varphi$ decides φ.

In other words, any $\mathcal{B}$-truth-tracking method is guaranteed to eventually find the correct truth value of φ, if worlds outside S_φ are deemed impossible.

Plurality coalitions. Requiring a strict majority of experts as above may be deemed too strong an assumption; we now consider a weaker notion. Write $\mathrm{Bel}_W(i, c) = \{\varphi \in \mathcal{L}_0 \mid W, c \models_{\mathcal{B}} \mathsf{B}_i\varphi\}$ for the beliefs of i in case c. Say $I \subseteq \mathcal{S}$ is a *coalition* (w.r.t W) if $\bigcup_{i \in I} \mathrm{Bel}_W(i, c)$ is consistent for each $c \in \mathcal{C}$. Note that the set of sources $J_W = \{i \mid \forall c \in \mathcal{C}, W, c \models \mathrm{Bel}_W(i, c)\}$ with correct beliefs forms a coalition, which we call the *correct coalition*.

Now let S_{pl} be the set of worlds in which the correct coalition forms a *plurality*, i.e. $W \in S_{\mathsf{pl}}$ iff $|J_W| > |I|$ for all coalitions I w.r.t W such that $I \neq J_W$. Note that J_W need not be a majority: we may have $|J_W| \leq \frac{|\mathcal{S}|}{2}$.

Then for any $W \in S_{\mathsf{pl}}$, and $c \in \mathcal{C}$, we have $Q^*_{\mathcal{B}}[W] \cap S_{\mathsf{pl}}, c \models \bigcup_{i \in J_W} \mathrm{Bel}_W(i, c)$. That is, any $\mathcal{B}$-truth-tracking method – when restricted to consider only worlds in S_{pl} – eventually comes to believe all the correct sources in J_W.

Indeed, if $W' \in Q^*_{\mathcal{B}}[W] \cap S_{\mathsf{pl}}$ then $\mathrm{Bel}_W(i, c) = \mathrm{Bel}_{W'}(i, c)$ for all i and c, so $I \subseteq \mathcal{S}$ is a coalition w.r.t W if and only if I is a coalition w.r.t W'. In particular, J_W is a coalition w.r.t W', which implies $|J_W| \leq |J_{W'}|$. Likewise, $J_{W'}$ is a coalition w.r.t W, so $|J_{W'}| \leq |J_W|$. Hence $|J_W| = |J_{W'}|$. Since J_W is the largest coalition w.r.t W, this means $J_W = J_{W'}$. Thus, recalling that $i \in J_{W'}$ implies $W', c \models \mathrm{Bel}_{W'}(i, c)$, we have $W', c \models \bigcup_{i \in J_W} \mathrm{Bel}_W(i, c)$ as required.

Example 6.6.3. *Consider the setting* $\mathsf{Prop} = \{p, q\}$, $\mathcal{S} = \{i_1, i_2, i_3, j_1, j_2\}$, $\mathcal{C} = \{c, d, e\}$. *One can construct* W *and* $\mathcal{B}$ *such that* $\neg p$ *holds in each case, and the beliefs of sources are described by the following sequence:*

$$\begin{aligned}\sigma = (&\langle i_1, c, p\rangle, \langle i_2, c, p \wedge q\rangle, \langle i_3, c, p \wedge \neg q\rangle, \langle j_1, c, \neg p\rangle, \langle j_2, c, \neg p\rangle,\\ &\langle i_1, d, p \wedge q\rangle, \langle i_2, d, p \wedge \neg q\rangle, \langle i_3, d, p\rangle, \langle j_1, d, \neg p\rangle, \langle j_2, d, \neg p\rangle,\\ &\langle i_1, e, p \wedge q\rangle, \langle i_2, e, p\rangle, \langle i_3, e, p \wedge \neg q\rangle, \langle j_1, e, \neg p\rangle, \langle j_2, e, \neg p\rangle).\end{aligned}$$

Observe that j_1 *and* j_2 *report* $\neg p$ *in each case, but are out-voted by* i_1, i_2 *and* i_3*, whose reports imply* p*. However,* i_2 *and* i_3 *disagree on* q *in case* c*,* i_1 *and* i_2 *disagree in case* d*, and* i_3 *disagree in case* e*. Consequently the only non-empty coalitions are* $\{i_1\}$*,* $\{i_2\}$*,* $\{i_3\}$ *and* $\{j_1, j_2\}$*, with* J_W *being the latter. We see that even though* j_1 *and* j_2 *are in the minority on* p *in each case, they form the largest coalition consistent* across cases*, and thus* $W \in S_{\mathsf{pl}}$*. In particular, any* $\mathcal{B}$*-truth-tracking method eventually believes* $\neg p$ *in all cases.*

Clearly these two choices of S are not exhaustive, and the reader may conceive of other reasonable assumptions to be placed on worlds in order to improve the informativeness of truth-tracking with belief streams. We leave a more in-depth study to future work; in the meantime it is hoped that the above results show the approach is at least feasible, and yields interesting results in the aforementioned special cases.

6.7 Conclusion

Summary. In this chapter we studied truth-tracking in the presence of non-expert sources. To start with, the model assumes sources report everything true *up to their lack of expertise*, i.e. all that they consider possible. We obtained precise characterisations of when truth-tracking methods can uniquely find the valuations of a world W, and in doing so showed how sources may *combine* their expertise to track the truth. Similar results were presented for finding the actual partitions of a world W, i.e. finding the true extent of each source's expertise.

We then presented the **Credulity** postulate, which characterises truth-tracking methods under mild assumptions. Roughly speaking, this postulate says that one needs to *trust* sources to be experts wherever possible. Purely deductive reasoning – in which one does not conjecture beyond the fact that each received report is sound – fails to be credulous enough in this sense, and thus some amount of *non-monotonic* reasoning is required for truth-tracking.

Next, we reconsidered the belief and expertise operators of Chapter 5 in the context of truth-tracking. Interestingly, it was seen that the variable-based conditioning method L_{vbc} is *not* truth-tracking, but the partition-based conditioning method L_{pbc} and score-based method L_{exm} are. The success of the latter two methods showed that truth-tracking is compatible with rational belief change as expressed by the postulates of Chapter 5 (which are satisfied by all three methods).

Finally, we introduced a mechanism by which sources form *beliefs* – possibly going beyond their knowledge, and possibly false in the actual world. Similar notions of truth-tracking were introduced, and some examples were given of how certain assumptions on worlds can improve the informativeness of truth-tracking in this setting.

Limitations and future work. Conceptually, the assumption that streams are complete is very strong. As seen in Example 6.2.1, completeness requires sources to give jointly inconsistent reports whenever $\Pi_i[v_c]$ contains more than just v_c. Such reports provide information about the source's expertise: if i reports both φ and $\neg\varphi$ we know $\neg\mathsf{E}_i\varphi$. To provide all sound reports sources must also have *negative introspection* over their own knowledge, i.e. they *know* when they do not know something. Indeed, as seen in Chapters 4 and 5, soundness is closely related to *S5 knowledge* (when partitions are used to represent expertise) which has been criticised in the philosophical literature as too strong.

While this was partially remedied in Section 6.6, the results there are somewhat preliminary. For instance, we did not proceed systematically in identifying assumptions under which belief-based truth-tracking can be made more informative, in the sense of revealing more information about the actual world W. Instead, we presented two ad-hoc examples based on the number of expert sources and coalitions of jointly consistent sources. It also remains to investigate our concrete example methods from Chapter 5 in relation to belief-based truth-tracking. Some of these methods were shown to be truth-tracking on sound and complete streams in Examples 6.5.2 and 6.5.4; are they also $\mathcal{B}$-truth-tracking?

On the technical side, our results on solvability of Q^* and the characterisation of Theorem 6.5.1 rely on the fact that we only consider finitely many worlds. In a sense this trivialises the problem of induction as studied by Kelly, Schulte, and Hendricks (1997) and Baltag, Gierasimczuk, and Smets (2016), among others. In future work it would be interesting to see which results can be carried over to the case where Prop is infinite.

— 7 —

Conclusion

Summary. This thesis has studied topics surrounding trustworthiness and expertise. Table 7.1 shows a rough overview of the different themes tackled in each chapter. We now reflect chapter-by-chapter on these themes in more detail, and discuss the connections between chapters.

In Chapter 2 we took a social choice perspective on truth discovery. Truth discovery concerns aggregating reports from unreliable sources, and assessing their trustworthiness in the process. The representation of information was simple in this chapter: we considered a number of "objects" of interest, each of which takes categorical values. Trustworthiness was expressed via a ranking of the sources, allowing us to say when one source should be considered *more trustworthy* than another. We would later take a different approach in Chapter 5, with information represented by propositional formulas and with expertise-based semantics for trustworthiness. However, the open-ended nature of ranking-based trustworthiness without precise semantics allowed us to define several truth discovery methods from the literature in our framework. Most importantly, Chapter 2 introduced several *axioms* for truth discovery. We proved the first impossibility results for truth discovery; this showed there are fundamental constraints on which properties one can hope for when constructing new truth discovery methods. The axioms highlighted a possible problem with Sums – a well known method introduced by Pasternack and Roth (2010) – in its failure of **Disjoint-independence**. This led us to define a new method, called UnboundedSums, which modifies Sums to resolve this axiomatic failure.

Table 7.1: Overview of the themes covered by each chapter of the thesis.

	TD	Tournaments	Modal	Belief change	Truth-tracking
Aggregating reports	✓			✓	✓
Assessing trustworthiness	✓	✓		✓	✓
Axiomatic analysis	✓	✓		✓	✓
Defining new methods	✓	✓		✓	
Reasoning with expertise			✓	✓	
Learning the truth					✓

Chapter 3 stayed with the social choice methodology, this time applied to bipartite tournament ranking. This problem was motivated by truth discovery – in particular, by semi-supervised truth discovery – and also relates to assessing trustworthiness of sources. Concretely, bipartite tournaments naturally model situations where the performance of sources on a number of "ground truth" objects is known in advance. One can then use bipartite tournament ranking methods to rank the sources by trustworthiness (with respect to the ground truth) and the objects by difficulty. But in fact, we gave several other example domains where bipartite tournaments provide a natural model, such as education and solo sports. We therefore took a more general view throughout the chapter and considered *abstract* bipartite tournaments, where no fixed interpretation was assigned to the players involved or the nature of the matches between them. The overall approach was again axiomatic. We introduced several axioms; some adaptations of standard social choice ideas and some specific to bipartite tournaments (e.g. **Dual**). We also paid special attention to the class of *chain-minimal* operators, the ranking operators associated with the graph modification problem of *chain editing*. Such operators were rationalised by a result showing them to be maximum likelihood estimators in a particular probabilistic model. In a sense this is a truth-tracking result: chain-minimal operators find the "true" strengths of the tournament participants, if one accepts the assumptions of the probabilistic model. However, this is a different sense of truth-tracking as studied in Chapter 6, which refers to finding true *information* when given conflicting reports.

The remainder of the thesis employed logic-based formalisms. The modal logic framework of Chapter 4 studied expertise and "soundness" of information. Here we took a deep dive into conceptual and mathematical properties of expertise. For instance, we established precise connections between expertise and *knowledge* via one-to-one mappings between classes of expertise models and relational models of epistemic logic. Roughly speaking, we saw that expertise on φ is equivalent to the statement that the source knows $\neg\varphi$ in all states where φ is false. The properties of knowledge involved in this translation depend on the properties one takes for expertise; if expertise collections are closed under intersections and unions we get S4 knowledge, and closure under complements strengthens this to S5. Technical results included axiomatisation results for various classes of expertise models, with completeness for the class of all models requiring a novel and non-trivial proof technique.

Importantly, Chapter 4 laid the groundwork for subsequent chapters. Chapter 5 took up the expertise framework in pursuit of roughly the same goals as truth discovery: to handle conflicting reports from multiple sources, and to both form beliefs about the world and assess the trustworthiness of sources. The *objects* of truth discovery played a similar role to the *cases* in the belief change problem; in both problems the patterns of reports across objects/cases allows operators to reason about the trustworthiness or expertise of sources. The representations in use differed, though: in Chapter 2 we used rankings of sources and claims, whereas Chapter 5 used sets of logical formulas. While rankings also played a major role in the construction of our belief and expertise operators, it was through rankings of *worlds* rather than sources. We argued on p. 107 that neither choice is better than the other; they merely give different perspectives on the broader problem. The result was that our axioms/postulates and operators share high level intuitions in places, but vary in their technical manifestation.

Finally, Chapter 6 took the same framework of the belief change problem to consider not just how to form *rational* beliefs on the basis of non-expert information, but how to find the *truth*. This concerned both finding the facts about the world in each case – i.e. reliably *aggregating* reports – and finding the true expertise of each source. We saw that when sources make all possible sounds reports, an operator must be somewhat *trusting*; this was expressed formally by the **Credulity** axiom. Consequently truth-tracking is inherently *non-monotonic*. We also found precise conditions under which the facts of the world can be found by truth-tracking methods, which gives an insight into the extent to which truth-tracking is possible at all with non-expert sources. Indeed, the running example of Chapter 6 served as an example where the truth cannot be found in some cases due to lack of relevant expertise.

Future Work. There are many directions for future work. Since specific open problems were discussed at the end of each chapter, in these final pages we reflect more generally on the thesis as a whole.

As we have often emphasised, Chapters 2 and 5 introduced new frameworks. While they were heavily inspired by existing literature (in social choice and belief change), both frameworks required significant research effort to develop, and their very formulation constituted a major portion of their respective chapters. This left less room to actually *use* these frameworks to obtain results. Indeed, the main contribution of these chapters was to lay the foundations, on which we hope future researchers may be able to build. In this sense we only scratched the surface of possible research questions. For example, with the axioms set out in Chapter 2 one could extend our analysis of Sums to further truth discovery methods (e.g. TruthFinder and CRH, which were defined in the framework but not studied in detail), consider axiomatic characterisations (e.g. axioms which characterise Sums), or consider truth-tracking in axiomatic terms. In Chapter 5 one could look to develop more sophisticated belief and expertise operators, explore computational issues (which we largely neglected), or consider extra postulates. These same considerations also apply to a lesser extent to Chapter 3, which introduced the new problem of bipartite tournament ranking.

This thesis has also been highly *theoretical* and even *mathematical* in its nature. Its contributions are formal definitions, theorems and proofs. It is hoped that the definitions are meaningful and the theorems interesting, but this is not enough to be useful in a *practical* sense. Building on the theoretical results with practical tools remains an important line of research. For example, the tractable chain-definable ranking methods introduced in Chapter 3 could be implemented and applied to real-world problems, such as truth discovery or the education example described on p. 59. One could work towards applying the belief and expertise operators of Chapter 5 to real data; the reliance a propositional logic representation presents a challenge here, but may lend itself well to domains with structured inputs (as opposed to natural language, say).

Bibliography

Alchourrón, Carlos E, Peter Gärdenfors, and David Makinson (1985). "On the logic of theory change: Partial meet contraction and revision functions". In: *Journal of symbolic logic*, pp. 510–530 (cited on pages 8, 107, 147, 148, 176, 202).

Altman, Alon and Moshe Tennenholtz (2005). "Ranking systems: the PageRank axioms". In: *Proceedings of the 6th ACM conference on Electronic commerce*. ACM, pp. 1–8 (cited on pages 2, 15, 55).

— (2008). "Axiomatic Foundations for Ranking Systems". In: *J. Artif. Int. Res.* 31.1, pp. 473–495. ISSN: 1076-9757. URL: http://dl.acm.org/citation.cfm?id=1622655.1622669 (cited on pages 15, 19, 26, 53, 54, 56).

Anđelić, Milica, Enide Andrade, Domingos M. Cardoso, Carlos M. da Fonseca, Slobodan K. Simić, and Dejan V. Tošić (2015). "Some new considerations about double nested graphs". In: *Linear Algebra and its Applications* 483, pp. 323–341 (cited on page 104).

Andersen, Reid, Christian Borgs, Jennifer Chayes, Uriel Feige, Abraham Flaxman, Adam Kalai, Vahab Mirrokni, and Moshe Tennenholtz (2008). "Trust-based Recommendation Systems: An Axiomatic Approach". In: *Proceedings of the 17th International Conference on World Wide Web*. WWW '08. New York, NY, USA: ACM, pp. 199–208. DOI: 10.1145/1367497.1367525. URL: http://doi.acm.org/10.1145/1367497.1367525 (cited on page 15).

Arrow, Kenneth J. (1952). "Social Choice and Individual Values". In: *Ethics* 62.3, pp. 220–222 (cited on pages 2, 35).

Axler, Sheldon (2014). *Linear Algebra Done Right*. Springer International Publishing. DOI: 10.1007/978-3-319-11080-6. URL: https://doi.org/10.1007/978-3-319-11080-6 (cited on page 250).

Baltag, Alexandru, Nina Gierasimczuk, Aybüke Özgün, Ana Lucia Vargas Sandoval, and Sonja Smets (2019). "A dynamic logic for learning theory". In: *Journal of Logical and Algebraic Methods in Programming* 109, p. 100485 (cited on page 202).

Baltag, Alexandru, Nina Gierasimczuk, and Sonja Smets (2016). "On the Solvability of Inductive Problems: A Study in Epistemic Topology". In: *Elec-*

tronic Proceedings in Theoretical Computer Science 215, pp. 81–98. ISSN: 2075-2180. DOI: 10.4204/eptcs.215.7. URL: http://dx.doi.org/10.4204/EPTCS.215.7 (cited on pages 11, 205, 207, 231).

Baltag, Alexandru, Nina Gierasimczuk, and Sonja Smets (2019). "Truth-Tracking by Belief Revision". In: *Studia Logica* 107.5, pp. 917–947. ISSN: 1572-8730. DOI: 10.1007/s11225-018-9812-x. URL: https://doi.org/10.1007/s11225-018-9812-x (cited on pages 1, 10, 11, 202, 203, 205).

Baltag, Alexandru and Bryan Renne (2016). "Dynamic Epistemic Logic". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2016. Metaphysics Research Lab, Stanford University (cited on page 6).

Baltag, Alexandru and Sonja Smets (2008). "A qualitative theory of dynamic interactive belief revision". In: *Logic and the foundations of game and decision theory (LOFT 7)* 3, pp. 9–58 (cited on pages 6, 142).

van Benthem, Johan and Guram Bezhanishvili (2007). "Modal logics of space". In: *Handbook of spatial logics*. Springer, pp. 217–298 (cited on pages 117, 130).

van Benthem, Johan, David Fernández-Duque, and Eric Pacuit (2012). "Evidence logic: A new look at neighborhood structures". In: *Advances in Modal Logic*. College Publications (cited on page 6).

— (2014). "Evidence and plausibility in neighborhood structures". In: *Annals of Pure and Applied Logic* 165.1, pp. 106–133 (cited on pages 6, 109).

van Benthem, Johan and Eric Pacuit (2011). "Dynamic logics of evidence-based beliefs". In: *Studia Logica* 99.1, pp. 61–92 (cited on pages 6, 109).

Berti-Equille, Laure and Javier Borge-Holthoefer (2015). "Veracity of data: From truth discovery computation algorithms to models of misinformation dynamics". In: *Synthesis Lectures on Data Management* 7.3, pp. 1–155 (cited on page 13).

Blackburn, Patrick, Maarten de Rijke, and Yde Venema (2001). *Modal Logic*. Cambridge University Press (cited on pages 122, 130).

Blum, Christian and Christina Isabel Zuber (2016). "Liquid democracy: Potentials, problems, and perspectives". In: *Journal of Political Philosophy* 24.2, pp. 162–182 (cited on page 108).

Booth, Richard (2005). "On the Logic of Iterated Non-prioritised Revision". In: *Conditionals, Information, and Inference*. Ed. by Gabriele Kern-Isberner, Wilhelm Rödder, and Friedhelm Kulmann. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 86–107 (cited on page 9).

Booth, Richard and Aaron Hunter (2018). "Trust as a Precursor to Belief Revision". In: *Journal of Artificial Intelligence Research* 61, pp. 699–722 (cited on pages 9, 149, 152, 191, 198–200).

Booth, Richard and Thomas Meyer (2011). "Belief Change". In: *Journal of Indian Council of Philosophical Research* 27.2, pp. 379–412 (cited on pages 1, 7).

Boutilier, Craig, Nir Friedman, and Joseph Y Halpern (1998). "Belief revision with unreliable observations". In: *AAAI/IAAI*, pp. 127–134 (cited on pages 164, 174).

Bouyssou, Denis (2004). “Monotonicity of ‘ranking by choosing’: A progress report”. In: *Social Choice and Welfare* 23.2, pp. 249–273 (cited on pages 87, 92).

Bozóki, Sándor, László Csató, and József Temesi (2016). “An application of incomplete pairwise comparison matrices for ranking top tennis players”. In: *European Journal of Operational Research* 248.1, pp. 211–218 (cited on page 58).

Bradley, Ralph Allan and Milton E. Terry (1952). “Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons”. In: *Biometrika* 39.3/4, pp. 324–345. ISSN: 00063444 (cited on pages 64, 65).

Brandt, Felix, Markus Brill, Paul Harrenstein, and Hervé Moulin (2016). “Tournament Solutions”. In: *Handbook of Computational Social Choice*. Ed. by Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel Procaccia. Cambridge University Press, pp. 57–84 (cited on pages 2–4, 58, 61, 63, 104).

Brandt, Felix, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel Procaccia, eds. (2016a). *Handbook of Computational Social Choice*. Cambridge University Press. DOI: 10.1017/CBO9781107446984 (cited on page 3).

— (2016b). “Introduction to Computational Social Choice”. In: *Handbook of Computational Social Choice*. Ed. by Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel Procaccia. 1st. New York, NY, USA: Cambridge University Press. Chap. 1 (cited on page 1).

Chi, Michelene TH, Robert Glaser, and Marshall J Farr (2014). *The nature of expertise*. Psychology Press (cited on page 108).

Christoff, Zoé and Davide Grossi (2017). “Binary Voting with Delegable Proxy: An Analysis of Liquid Democracy”. In: *TARK* (cited on pages 18, 36).

Collins, Harry and Robert Evans (2008). *Rethinking expertise*. University of Chicago Press (cited on page 108).

Condorcet, Nicolas de (1785). *Essai sur l’application de l’analyse à la probabilité des décisions rendues à la pluralité des voix* (cited on page 201).

Conitzer, Vincent and Toby Walsh (2016). “Barriers to Manipulation in Voting”. In: *Handbook of Computational Social Choice*. Ed. by Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel Procaccia. 1st. New York, NY, USA: Cambridge University Press. Chap. 6 (cited on page 3).

Cook II, David (2015). “Nested colorings of graphs”. In: *Australasian Journal of Combinatorics* 62.1, pp. 100–127 (cited on page 104).

Csató, László (2019). “An impossibility theorem for paired comparisons”. In: *Central European Journal of Operations Research* 27.2, pp. 497–514 (cited on pages 59, 60).

Darwiche, Adnan and Judea Pearl (1997). “On the logic of iterated belief revision”. In: *Artificial Intelligence* 89.1-2, pp. 1–29. DOI: 10.1016/s0004-3702(96)00038-0. URL: https://doi.org/10.1016%2Fs0004-3702%2896%2900038-0 (cited on page 8).

Dastani, Mehdi, Andreas Herzig, Joris Hulstijn, and Leendert van der Torre (2004). "Inferring trust". In: *International Workshop on Computational Logic in Multi-Agent Systems*. Springer, pp. 144–160 (cited on page 6).

Delgrande, James P, Didier Dubois, and Jérôme Lang (2006). "Iterated Revision as Prioritized Merging." In: *KR* 6, pp. 210–220 (cited on pages 159, 199, 200).

Delgrande, James P, Pavlos Peppas, and Stefan Woltran (2018). "General belief revision". In: *Journal of the ACM (JACM)* 65.5, pp. 1–34 (cited on pages 157, 165, 166).

Denny, Paul, John Hamer, Andrew Luxton-Reilly, and Helen Purchase (2008). "PeerWise: Students Sharing Their Multiple Choice Questions". In: *Proceedings of the Fourth International Workshop on Computing Education Research*. ICER '08. Sydney, Australia: ACM, pp. 51–58 (cited on page 59).

van Dijck, José and Donya Alinejad (2020). "Social Media and Trust in Scientific Expertise: Debating the Covid-19 Pandemic in The Netherlands". In: *Social Media + Society* 6.4 (cited on page 108).

Ding, Hu, Jing Gao, and Jinhui Xu (2016). "Finding global optimum for truth discovery: Entropy based geometric variance". In: *Proc. 32nd International Symposium on Computational Geometry (SoCG 2016)* (cited on page 19).

van Ditmarsch, Hans, Wiebe van der Hoek, and Barteld Kooi (2008). *Dynamic Epistemic Logic*. Springer Netherlands. DOI: 10.1007/978-1-4020-5839-4 (cited on pages 6, 109, 143).

Dokow, Elad and Ron Holzman (2010). "Aggregation of binary evaluations with abstentions". In: *Journal of Economic Theory* 145, pp. 544–561 (cited on page 18).

Dong, Xin Luna, Laure Berti-Equille, and Divesh Srivastava (2009). "Truth Discovery and Copying Detection in a Dynamic World". In: *Proc. VLDB Endow.* 2.1, pp. 562–573. ISSN: 2150-8097. DOI: 10.14778/1687627.1687691. URL: https://doi.org/10.14778/1687627.1687691 (cited on pages 15, 16).

Drange, Pål Grønås, Markus Sortland Dregi, Daniel Lokshtanov, and Blair D. Sullivan (2015). "On the threshold of intractability". In: *Algorithms-ESA 2015*. Springer, pp. 411–423 (cited on pages 64, 103).

Du, Yang, Yu-E Sun, He Huang, Liusheng Huang, Hongli Xu, Yu Bao, and Hansong Guo (2019). "Bayesian Co-Clustering Truth Discovery for Mobile Crowd Sensing Systems". In: *IEEE Transactions on Industrial Informatics*, pp. 1–1. ISSN: 1551-3203. DOI: 10.1109/TII.2019.2896287 (cited on pages 13, 22).

Elkind, Edith and Arkadii Slinko (2016). "Rationalizations of voting rules". In: *Handbook of Computational Social Choice*. Ed. by Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel Procaccia. Cambridge University Press, pp. 169–196 (cited on pages 64, 65, 104, 201).

Endriss, Ulle (2016). "Judgment Aggregation". In: *Handbook of Computational Social Choice*. Ed. by Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme

Lang, and Ariel Procaccia. 1st. New York, NY, USA: Cambridge University Press. Chap. 17 (cited on pages 2, 15, 18, 35).

Endriss, Ulle (2020). "Analysis of one-to-one matching mechanisms via SAT solving: Impossibilities for universal axioms". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 02, pp. 1918–1925 (cited on page 3).

Ericsson, K Anders and Tyler J Towne (2010). "Expertise." In: *WIREs Cognitive Science* (cited on page 108).

Everaere, Patricia, Sébastien Konieczny, Pierre Marquis, et al. (2010). "The Epistemic View of Belief Merging: Can We Track the Truth?." In: *ECAI*, pp. 621–626 (cited on page 201).

Everaere, Patricia, Sébastien Konieczny, and Pierre Marquis (2020). "Belief Merging Operators as Maximum Likelihood Estimators". In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. Ed. by Christian Bessiere. International Joint Conferences on Artificial Intelligence Organization, pp. 1763–1769 (cited on page 64).

Fagin, Ronald, Yoram Moses, Joseph Y Halpern, and Moshe Y Vardi (2003). *Reasoning about knowledge*. MIT press (cited on pages 109, 134, 135, 137).

Fermé, Eduardo and Sven Ove Hansson (1999). "Selective revision". In: *Studia Logica* 63.3, pp. 331–342 (cited on pages 9, 149, 184, 191, 199, 200).

— (2018). *Belief Change*. SpringerBriefs in Intelligent Systems. Springer. DOI: `https://doi.org/10.1007/978-3-319-60535-7` (cited on pages 1, 7).

Galland, Alban, Serge Abiteboul, Amélie Marian, and Pierre Senellart (2010). "Corroborating Information from Disagreeing Views". In: *Proceedings of the Third ACM International Conference on Web Search and Data Mining*. WSDM '10. New York, New York, USA: ACM, pp. 131–140. DOI: `10.1145/1718487.1718504`. URL: `http://doi.acm.org/10.1145/1718487.1718504` (cited on pages 14, 19, 22, 59).

Garson, James (2021). "Modal Logic". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University (cited on page 5).

Ghorbani, Ebrahim (2017). *Some spectral properties of chain graphs*. arXiv: `1703.03581 [math.CO]` (cited on page 104).

Ghosh, Arpita, Satyen Kale, and Preston McAfee (2011). "Who moderates the moderators?" In: *Proceedings of the 12th ACM conference on Electronic commerce - EC 11*. ACM Press. DOI: `10.1145/1993574.1993599`. URL: `https://doi.org/10.1145%2F1993574.1993599` (cited on page 14).

Gierasimczuk, Nina (2009). "Bridging learning theory and dynamic epistemic logic". In: *Synthese* 169.2, pp. 371–384 (cited on page 6).

— (2010). "Knowing one's limits: logical analysis of inductive inference". PhD thesis. Institute for Logic, Language and Computation, University of Amsterdam (cited on pages 1, 6, 10).

Gold, E Mark (1967). "Language identification in the limit". In: *Information and Control* 10.5, pp. 447–474. DOI: `10.1016/s0019-9958(67)91165-5` (cited on page 201).

Goldman, Alvin I (2018). "Expertise". In: *Topoi* 37.1, pp. 3–10 (cited on page 108).

González-Díaz, Julio, Ruud Hendrickx, and Edwin Lohmann (2014). "Paired comparisons analysis: an axiomatic approach to ranking methods". In: *Social Choice and Welfare* 42.1, pp. 139–169 (cited on pages 4, 58–60, 64, 65, 71, 104).

Goranko, Valentin and Solomon Passy (1992). "Using the Universal Modality: Gains and Questions". In: *Journal of Logic and Computation* 2.1, pp. 5–30 (cited on page 110).

Goranko, Valentin and Antje Rumberg (2022). "Temporal Logic". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2022. Metaphysics Research Lab, Stanford University (cited on page 5).

Grégoire, Eric and Sébastien Konieczny (2006). "Logic-based approaches to information fusion". In: *Information Fusion* 7.1. Logic-based Approaches to Information Fusion, pp. 4–18. ISSN: 1566-2535. DOI: `https://doi.org/10.1016/j.inffus.2005.08.001`. URL: `http://www.sciencedirect.com/science/article/pii/S1566253505000771` (cited on page 197).

Grofman, Bernard, Guillermo Owen, and Scott L Feld (1983). "Thirteen theorems in search of the truth". In: *Theory and decision* 15.3, pp. 261–278 (cited on pages 201, 228).

Grove, Adam (1988). "Two modellings for theory change". In: *Journal of philosophical logic*, pp. 157–170 (cited on page 176).

Gupta, Manish and Jiawei Han (2011). "Heterogeneous Network-based Trust Analysis: A Survey". In: *SIGKDD Explor. Newsl.* 13.1, pp. 54–71. ISSN: 1931-0145. DOI: `10.1145/2031331.2031341`. URL: `http://doi.acm.org/10.1145/2031331.2031341` (cited on pages 13, 14).

Hansson, Sven Ove (1999a). "A survey of non-prioritized belief revision". In: *Erkenntnis* 50.2, pp. 413–427 (cited on pages 9, 148).

— (1999b). *A Textbook of Belief Dynamics. Theory Change and Database Updating*. Kluwer Academic Publishers (cited on page 8).

— (2022). "Logic of Belief Revision". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2022. Metaphysics Research Lab, Stanford University (cited on pages 1, 7).

Hansson, Sven Ove, Eduardo Fermé, John Cantwell, and Marcelo Alejandro Falappa (2001). "Credibility limited revision". In: *Journal of Symbolic Logic* 66.4, pp. 1581–1596. DOI: `10.2307/2694963`. URL: `https://doi.org/10.2307%2F2694963` (cited on pages 9, 149, 199, 200).

Hartmann, Stephan and Jan Sprenger (2012). "Judgment aggregation and the problem of tracking the truth". In: *Synthese* 187.1, pp. 209–221. ISSN: 1573-0964. DOI: `10.1007/s11229-011-0031-5`. URL: `https://doi.org/10.1007/s11229-011-0031-5` (cited on page 201).

Herzig, Andreas, Emiliano Lorini, Jomi F Hübner, and Laurent Vercouter (2010). "A logic of trust and reputation". In: *Logic Journal of the IGPL* 18.1, pp. 214–244 (cited on page 6).

Holliday, Wesley H (2009). "Dynamic testimonial logic". In: *International Workshop on Logic, Rationality and Interaction*. Springer, pp. 161–179 (cited on page 6).

Hunter, Aaron (2021). "Building Trust for Belief Revision". In: *PRICAI 2021: Trends in Artificial Intelligence*. Ed. by Duc Nghia Pham, Thanaruk Theeramunkong, Guido Governatori, and Fenrong Liu. Cham: Springer International Publishing, pp. 543–555 (cited on pages 150, 198–200).

Jain, Sanjay, Daniel Osherson, James S Royer, Arun Sharma, et al. (1999). *Systems that learn: an introduction to learning theory*. MIT (cited on pages 1, 10, 201, 202).

Jiao, Yang, R Ravi, and Wolfgang Gatterbauer (2017). "Algorithms for automatic ranking of participants and tasks in an anonymized contest". In: *International Workshop on Algorithms and Computation*. Springer, pp. 335–346 (cited on pages 5, 59, 60, 63, 69, 103, 105).

Katsuno, Hirofumi and Alberto O. Mendelzon (1991). "Propositional knowledge base revision and minimal change". In: *Artificial Intelligence* 52.3, pp. 263–294. DOI: 10.1016/0004-3702(91)90069-v. URL: https://doi.org/10.1016%2F0004-3702%2891%2990069-v (cited on page 176).

— (1992). "On the difference between updating a knowledge base and revising it". In: *Belief Revision*. Ed. by Peter Gärdenfors. Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, pp. 183–203. DOI: 10.1017/CBO9780511526664.007 (cited on page 8).

Kelly, Kevin (1998). "The learning power of belief revision". In: *TARK*. Vol. 98. Citeseer, pp. 111–124 (cited on page 10).

— (1999). "Iterated Belief Revision, Reliability, and Inductive Amnesia". In: *Erkenntnis* 50.1, pp. 7–53. DOI: 10.1023/a:1005444112348. URL: https://doi.org/10.1023%2Fa%3A1005444112348 (cited on page 10).

Kelly, Kevin, Oliver Schulte, and Vincent Hendricks (1997). "Reliable belief revision". In: *Logic and Scientific Methods*. Springer, pp. 383–398 (cited on pages 10, 202, 231).

Kilov, Daniel (2021). "The brittleness of expertise and why it matters". In: *Synthese* 199.1, pp. 3431–3455 (cited on page 108).

Kleinberg, Jon M. (1999). "Authoritative Sources in a Hyperlinked Environment". In: *J. ACM* 46.5, pp. 604–632. ISSN: 0004-5411. DOI: 10.1145/324133.324140. URL: http://doi.acm.org/10.1145/324133.324140 (cited on pages 22, 46, 55).

Konieczny, Sébastien and Ramón Pino Pérez (2002). "Merging information under constraints: a logical framework". In: *Journal of Logic and Computation* 12.5, pp. 773–808 (cited on pages 8, 10, 107, 147, 149, 196, 197, 202).

— (2011). "Logic Based Merging". In: *Journal of Philosophical Logic* 40.2, pp. 239–270. DOI: 10.1007/s10992-011-9175-5. URL: https://doi.org/10.1007%2Fs10992-011-9175-5 (cited on page 196).

Kotonya, Neema and Francesca Toni (2020). "Explainable Automated Fact-Checking for Public Health Claims". In: *Proceedings of the 2020 Conference* *on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7740–7754 (cited on page 14).

Kraus, Sarit, Daniel Lehmann, and Menachem Magidor (1990). "Nonmonotonic reasoning, preferential models and cumulative logics". In: *Artificial intelligence* 44.1-2, pp. 167–207 (cited on page 165).

Kruger, Justin, Ulle Endriss, Raquel Fernandez, and Ciyang Qing (2014). "Axiomatic Analysis of Aggregation Methods for Collective Annotation". In: *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*. AAMAS '14. Paris, France: International Foundation for Autonomous Agents and Multiagent Systems, pp. 1185–1192. URL: http://dl.acm.org/citation.cfm?id=2617388.2617437 (cited on page 2).

Laslier, Jean-François and M Remzi Sanver (2010). *Handbook On Approval Voting*. Springer Science & Business Media (cited on page 53).

Li, Yaliang, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han (2016). "A Survey on Truth Discovery". In: *SIGKDD Explorations Newsletter* 17.2, pp. 1–16. ISSN: 1931-0145. DOI: 10.1145/2897350.2897352. URL: http://doi.acm.org/10.1145/2897350.2897352 (cited on pages 3, 13, 20, 201).

Li, Yaliang, Qi Li, Jing Gao, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han (2016). "Conflicts to Harmony: A Framework for Resolving Conflicts in Heterogeneous Data by Truth Discovery". In: *IEEE Transactions on Knowledge and Data Engineering* 28.8, pp. 1986–1999. ISSN: 1041-4347. DOI: 10.1109/TKDE.2016.2559481 (cited on pages 14–16, 19, 20, 22, 24).

Liau, Churn-Jung (2003). "Belief, information acquisition, and trust in multi-agent systems – A modal logic formulation". In: *Artificial Intelligence* 149.1, pp. 31–60 (cited on page 6).

Llewellyn, Sue (2020). "Covid-19: how to be careful with trust and expertise on social media". In: *BMJ* 368 (cited on page 108).

Ma, Fenglong, Yaliang Li, Qi Li, Minghui Qiu, Jing Gao, Shi Zhi, Lu Su, Bo Zhao, Heng Ji, and Jiawei Han (2015). "FaitCrowd: Fine Grained Truth Discovery for Crowdsourced Data Aggregation". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '15. event-place: Sydney, NSW, Australia. New York, NY, USA: ACM, pp. 745–754. DOI: 10.1145/2783258.2783314. URL: http://doi.acm.org/10.1145/2783258.2783314 (cited on page 13).

Ma, Fenglong, Chuishi Meng, Houping Xiao, Qi Li, Jing Gao, Lu Su, and Aidong Zhang (2017). "Unsupervised Discovery of Drug Side-Effects from Heterogeneous Data Sources". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '17. Halifax, NS, Canada: ACM, pp. 967–976. DOI: 10.1145/3097983.3098129. URL: http://doi.acm.org/10.1145/3097983.3098129 (cited on page 13).

Makinson, David (1997). "Screened revision". In: *Theoria* 63.1-2, pp. 14–23 (cited on page 9).

Marshall, Jermaine, Arturo Argueta, and Dong Wang (2017). "A neural network approach for truth discovery in social sensing". In: *2017 IEEE 14th international conference on mobile Ad Hoc and sensor systems (MASS)*. IEEE, pp. 343–347 (cited on page 14).

May, Kenneth O (1952). "A set of independent necessary and sufficient conditions for simple majority decision". In: *Econometrica: Journal of the Econometric Society*, pp. 680–684 (cited on page 32).

McKinsey, J. C. C. (1941). "A Solution of the Decision Problem for the Lewis systems S2 and S4, with an Application to Topology". In: *The Journal of Symbolic Logic* 6.4, pp. 117–134. URL: http://www.jstor.org/stable/2267105 (visited on 09/29/2022) (cited on page 5).

McKinsey, J. C. C. and Alfred Tarski (1944). "The Algebra of Topology". In: *Annals of Mathematics* 45.1, pp. 141–191. URL: http://www.jstor.org/stable/1969080 (visited on 09/29/2022) (cited on page 5).

McNamara, Paul and Frederik Van De Putte (2022). "Deontic Logic". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Fall 2022. Metaphysics Research Lab, Stanford University (cited on page 5).

Meir, Reshef, Ofra Amir, Gal Cohensius, Omer Ben-Porat, and Lirong Xia (2019). "Truth Discovery via Proxy Voting". In: *CoRR* abs/1905.00629. arXiv: 1905.00629. URL: http://arxiv.org/abs/1905.00629 (cited on page 57).

Montague, Richard (1970). "Universal grammar". In: *Theoria* 36.3, pp. 373–398 (cited on pages 5, 111).

Moss, Lawrence S and Rohit Parikh (1992). "Topological Reasoning and The Logic of Knowledge." In: *TARK*. Vol. 92, pp. 95–105 (cited on page 6).

Natanzon, Assaf, Ron Shamir, and Roded Sharan (2000). "A polynomial approximation algorithm for the minimum fill-in problem". In: *SIAM Journal on Computing* 30.4, pp. 1067–1079 (cited on page 105).

— (2001). "Complexity classification of some edge modification problems". In: *Discrete Applied Mathematics* 113.1, pp. 109–128 (cited on page 103).

Olsson, Catherine, Surya Bhupatiraju, Tom Brown, Augustus Odena, and Ian Goodfellow (2018). *Skill Rating for Generative Models*. arXiv: 1808.04888 [stat.ML] (cited on page 59).

Özgün, Aybüke (2017). "Evidence in Epistemic Logic : A Topological Perspective". Theses. University of Amsterdam (cited on pages 5, 6, 117).

Pacuit, Eric (2017). *Neighborhood semantics for modal logic*. Springer International Publishing (cited on pages 5, 111).

Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd (1999). *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report 1999-66. Stanford InfoLab. URL: http://ilpubs.stanford.edu:8090/422/ (cited on page 55).

Pasternack, Jeff and Dan Roth (2010). "Knowing What to Believe (when You Already Know Something)". In: *Proceedings of the 23rd International Conference on Computational Linguistics*. COLING '10. Beijing, China: Asso-

ciation for Computational Linguistics, pp. 877–885. URL: http://dl.acm.org/citation.cfm?id=1873781.1873880 (cited on pages 14–16, 19, 20, 22, 46, 232).
Plaza, Jan (2007). "Logics of public communications". In: *Synthese* 158.2, pp. 165–179 (cited on pages 6, 143).
Rekatsinas, Theodoros, Manas Joglekar, Hector Garcia-Molina, Aditya Parameswaran, and Christopher Ré (2017). "SLiMFast: Guaranteed results for data fusion and source reliability". In: *Proceedings of the 2017 ACM International Conference on Management of Data*, pp. 1399–1414 (cited on pages 4, 59, 149).
Rendsvig, Rasmus and John Symons (2021). "Epistemic Logic". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University (cited on page 5).
Rodenhäuser, Leif Benjamin (2014). "A matter of trust: Dynamic attitudes in epistemic logic". PhD thesis. Institute for Logic, Language and Computation, University of Amsterdam (cited on page 6).
Rubinstein, Ariel (1980). "Ranking the participants in a tournament". In: *SIAM Journal on Applied Mathematics* 38.1, pp. 108–111 (cited on pages 61, 70, 71).
Sakama, Chiaki, Martin Caminada, and Andreas Herzig (2014). "A Formal Account of Dishonesty". In: *Logic Journal of the IGPL* 23.2, pp. 259–294. ISSN: 1367-0751. DOI: 10.1093/jigpal/jzu043. URL: https://doi.org/10.1093/jigpal/jzu043 (cited on page 6).
Schwind, Nicolas and Sébastien Konieczny (2020). "Non-Prioritized Iterated Revision: Improvement via Incremental Belief Merging". In: *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning*, pp. 738–747. DOI: 10.24963/kr.2020/76. URL: https://doi.org/10.24963/kr.2020/76 (cited on page 159).
Scott, Dana (1970). "Advice on Modal Logic". In: *Philosophical Problems in Logic: Some Recent Developments*. Ed. by Karel Lambert. Dordrecht: Springer Netherlands, pp. 143–173. DOI: 10.1007/978-94-010-3272-8_7 (cited on pages 5, 111).
Singleton, Joseph (2021). "A Logic of Expertise". In: *ESSLLI 2021 Student Session*. URL: https://arxiv.org/abs/2107.10832 (cited on pages vii, 110).
Singleton, Joseph and Richard Booth (2020). "An Axiomatic Approach to Truth Discovery". In: *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '20. Virtual Event: International Foundation for Autonomous Agents and Multiagent Systems, pp. 2011–2013. URL: https://www.ifaamas.org/Proceedings/aamas2020/pdfs/p2011.pdf (cited on pages vii, 16).
— (2021). "Rankings for Bipartite Tournaments via Chain Editing". In: *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '21. Virtual Event: International Foundation for Autonomous Agents and Multiagent Systems, pp. 1236–1244. URL:

https://www.ifaamas.org/Proceedings/aamas2021/pdfs/p1236.pdf (cited on pages vii, 60).

Singleton, Joseph and Richard Booth (2022a). "Towards an axiomatic approach to truth discovery". In: *Autonomous Agents and Multi-Agent Systems* 36.2, pp. 1–49. URL: https://doi.org/10.1007/s10458-022-09569-3 (cited on pages vii, 16).

— (2022b). "Who's the Expert? On Multi-source Belief Change". In: *Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning*, pp. 331–340. URL: https://doi.org/10.24963/kr.2022/33 (cited on pages viii, 150).

— (2022c). "Truth-Tracking with Non-Expert Information Sources". In: *Proceedings of the 20th International Workshop on Non-Monotonic Reasoning*. NMR '22. CEUR Workshop Proceedings, pp. 80–91. URL: http://ceur-ws.org/Vol-3197/paper8.pdf (cited on pages viii, 203).

— (2023). "Expertise and information: an epistemic logic perspective". In: *Synthese* 201.2. URL: https://doi.org/10.1007/s11229-023-04064-y (cited on pages vii, 110).

Slutzki, Giora and Oscar Volij (2005). "Ranking participants in generalized tournaments". In: *International Journal of Game Theory* 33.2, pp. 255–270 (cited on pages 59, 60, 105).

— (2006). "Scoring of web pages and tournaments—axiomatizations". In: *Social Choice and Welfare* 26.1, pp. 75–92 (cited on page 58).

Spohn, Wolfgang (1988). "Ordinal conditional functions: A dynamic theory of epistemic states". In: *Causation in decision, belief change, and statistics*. Springer, pp. 105–134 (cited on page 164).

Steiner, Anne K (1966). "The lattice of topologies: structure and complementation". In: *Transactions of the American Mathematical Society* 122.2, pp. 379–398 (cited on page 136).

Tagliaferri, Mirko (2019). "A logical language for computational trust". PhD thesis. The University of Urbino (cited on page 6).

Tennenholtz, Moshe (2004). "Reputation Systems: An Axiomatic Approach". In: *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. UAI '04. Banff, Canada: AUAI Press, pp. 544–551. URL: http://dl.acm.org/citation.cfm?id=1036843.1036909 (cited on pages 2, 15, 56).

Terzopoulou, Zoi and Ulle Endriss (2019). "Optimal Truth-Tracking Rules for the Aggregation of Incomplete Judgments". In: *Proceedings of the 12th International Symposium on Algorithmic Game Theory (SAGT-2019)* (cited on page 201).

Waguih, Dalia Attia and Laure Berti-Equille (2014). "Truth discovery algorithms: An experimental evaluation". In: *arXiv preprint arXiv:1409.6428* (cited on page 14).

Wang, Dong, Lance Kaplan, Hieu Le, and Tarek Abdelzaher (2012). "On Truth Discovery in Social Sensing: A Maximum Likelihood Estimation Approach". In: *Proceedings of the 11th International Conference on Information Pro-*

cessing in Sensor Networks. IPSN '12. event-place: Beijing, China. New York, NY, USA: ACM, pp. 233–244. DOI: 10.1145/2185677.2185737. URL: http://doi.acm.org/10.1145/2185677.2185737 (cited on pages 13, 16, 64).

Wang, Yaqing, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao (2018). "Eann: Event adversarial neural networks for multi-modal fake news detection". In: *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pp. 849–857 (cited on page 14).

Wąs, Tomasz and Oskar Skibski (2020). *Axiomatic Characterization of PageRank*. DOI: 10.48550/ARXIV.2010.08487. URL: https://arxiv.org/abs/2010.08487 (cited on page 55).

Whyte, Kyle Powys and Robert P Crease (2010). "Trust, expertise, and the philosophy of science". In: *Synthese* 177.3, pp. 411–425 (cited on page 108).

Xaudiera, Sergi and Ana S Cardenal (2020). "Ibuprofen narratives in five European countries during the COVID-19 pandemic". In: *Harvard Kennedy School Misinformation Review* 1.3. URL: https://misinforeview.hks.harvard.edu/article/ibuprofen-narratives-in-five-european-countries-during-the-covid-19-pandemic/ (cited on page 108).

Xiao, Houping (2018). "Multi-sourced Information Trustworthiness Analysis: Applications and Theory". PhD thesis. University at Buffalo, State University of New York (cited on page 14).

Xiao, Houping, Jing Gao, Zhaoran Wang, Shiyu Wang, Lu Su, and Han Liu (2016). "A Truth Discovery Approach with Theoretical Guarantee". In: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, pp. 1925–1934. DOI: 10.1145/2939672.2939816. URL: http://doi.acm.org/10.1145/2939672.2939816 (cited on pages 14, 16).

Xiao, Houping and Shiyu Wang (2015). "A Joint Maximum Likelihood Estimation Framework for Truth Discovery: A Unified Perspective". In: *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1. DOI: 10.1109/TKDE.2022.3173911 (cited on page 20).

Yang, Yi, Quan Bai, and Qing Liu (2018). "On the Discovery of Continuous Truth: A Semi-supervised Approach with Partial Ground Truths". In: *Web Information Systems Engineering – WISE 2018*. Springer International Publishing, pp. 424–438. DOI: 10.1007/978-3-030-02922-7_29. URL: https://doi.org/10.1007%2F978-3-030-02922-7_29 (cited on page 19).

— (2019). "A probabilistic model for truth discovery with object correlations". In: *Knowledge-Based Systems* 165, pp. 360–373. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2018.12.004. URL: http://www.sciencedirect.com/science/article/pii/S0950705118305914 (cited on pages 15, 22, 56).

Yannakakis, Mihalis (1981). "Computing the minimum fill-in is NP-complete". In: *SIAM Journal on Algebraic Discrete Methods* 2.1, pp. 77–79 (cited on pages 5, 61, 103).

Yasser, Ammar and Haythem O. Ismail (2020). "Information Revision: The Joint Revision of Belief and Trust". In: *Proceedings of the 18th International Workshop on Non-Monotonic Reasoning*. NMR '20, pp. 150–159. URL: https://nmr2020.dc.uba.ar/WorkshopNotes.pdf (cited on pages 10, 149, 197, 200).

— (2021). "Trust Is All You Need: From Belief Revision to Information Revision". In: *Logics in Artificial Intelligence*. Ed. by Wolfgang Faber, Gerhard Friedrich, Martin Gebser, and Michael Morak. Cham: Springer International Publishing, pp. 50–65 (cited on page 197).

Yin, Xiaoxin, Jiawei Han, and Philip S. Yu (2008). "Truth Discovery with Multiple Conflicting Information Providers on the Web". In: *IEEE Transactions on Knowledge and Data Engineering* 20.6, pp. 796–808. ISSN: 1041-4347. DOI: 10.1109/TKDE.2007.190745 (cited on pages 14–16, 19, 20, 22–24).

Yin, Xiaoxin and Wenzhao Tan (2011). "Semi-supervised Truth Discovery". In: *Proceedings of the 20th International Conference on World Wide Web*. WWW '11. event-place: Hyderabad, India. New York, NY, USA: ACM, pp. 217–226. DOI: 10.1145/1963405.1963439. URL: http://doi.acm.org/10.1145/1963405.1963439 (cited on pages 4, 14, 59, 149).

Zach, Richard (2019). *Boxes and Diamonds: An Open Introduction to Modal Logic*. Open Logic Project (cited on page 116).

Zhang, Daniel Yue, Rungang Han, Dong Wang, and Chao Huang (2016). "On robust truth discovery in sparse social media sensing". In: *2016 IEEE International Conference on Big Data (Big Data)*, pp. 1076–1081. DOI: 10.1109/BigData.2016.7840710 (cited on pages 13, 19).

Zhang, Liyan, Guo-Jun Qi, Dong Zhang, and Jinhui Tang (2018). "Latent Dirichlet Truth Discovery: Separating Trustworthy and Untrustworthy Components in Data Sources". In: *IEEE Access* 6, pp. 1741–1752. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2017.2780182 (cited on pages 16, 19, 22).

Zhi, Shi, Bo Zhao, Wenzhu Tong, Jing Gao, Dian Yu, Heng Ji, and Jiawei Han (2015). "Modeling Truth Existence in Truth Discovery". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '15. Sydney, NSW, Australia: ACM, pp. 1543–1552. DOI: 10.1145/2783258.2783339. URL: http://doi.acm.org/10.1145/2783258.2783339 (cited on pages 19, 22).

Zwicker, William S. (2016). "Introduction to the Theory of Voting". In: *Handbook of Computational Social Choice*. Ed. by Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel Procaccia. 1st. New York, NY, USA: Cambridge University Press. Chap. 2 (cited on pages 3, 15, 26, 31, 73).

— A —

Proofs for Chapter 2

A.1 Proof of Theorem 2.4.4

Theorem 2.4.4. *UnboundedSums is ordinally convergent in the following sense: for every network N there is $m \in \mathbb{N}$ such that for all $n \geq m$, $s, s' \in S$ and $c, c' \in C$,*

$$\begin{aligned} T_N^n(s) \leq T_N^n(s') &\iff T_N^m(s) \leq T_N^m(s'), \\ T_N^n(c) \leq T_N^n(c') &\iff T_N^m(c) \leq T_N^m(c'). \end{aligned}$$

That is, the rankings induced by T_N^n are constant for $n \geq m$.

Proof. The proof will use some results from linear algebra, so we will work with a matrix and vector representation of UnboundedSums. Take some network N. Enumerate S as $\{s_1, \ldots, s_k\}$, and C as $\{c_1, \ldots, c_l\}$. Write M for the $k \times l$ matrix given by

$$[M]_{ij} = \begin{cases} 1 & \text{if } s_i \in \mathrm{src}_N(c_j), \\ 0 & \text{otherwise.} \end{cases} \qquad (1 \leq i \leq k, 1 \leq j \leq l)$$

Write v_n and w_n for the vectors of source and claims scores of UnboundedSums at iteration n, i.e.

$$v_n = [T_N^n(s_1), \ldots, T_N^n(s_k)]^\top \in \mathbb{R}^k,$$

$$w_n = [T_N^n(c_1), \ldots, T_N^n(c_l)]^\top \in \mathbb{R}^l.$$

Multiplication by M encodes the update step of UnboundedSums: it is easily shown that $v_{n+1} = Mw_n$ and $w_{n+1} = M^\top v_{n+1}$. Writing $A = MM^\top \in \mathbb{R}^{k \times k}$, we have $v_{n+1} = Av_n$, and therefore $v_{n+1} = A^n v_1$.

To show that the rankings of UnboundedSums remain constant after finitely many iterations, we will show that for each $s_p, s_q \in S$ there is $m_{pq} \in \mathbb{N}$ such that $\mathrm{sign}([v_n]_p - [v_n]_q)$ is constant for all $n \geq m_{pq}$. Since $[v_n]_p$ and $[v_n]_q$ are the trust scores of s_p and s_q respectively in the n-th iteration, this will show that the ranking of s_p and s_q remains the same after m_{pq} iterations. Since there are only finitely many pairs of sources, we may then take m as the maximum value of m_{pq} over all pairs (p, q), and the entire source ranking $\sqsubseteq_N^{T^n}$ of UnboundedSums remains constant for $n \geq m$. An almost identical argument can be carried out for the claim ranking, and these together will prove the result.

So, fix $s_p, s_q \in S$. Write $\delta_n = [v_n]_p - [v_n]_q$. First note that $A = MM^\top$ is symmetric, so the *spectral theorem* gives the existence of k orthogonal eigenvectors $x_1, \dots, x_k$ for A (Axler 2014, Theorem 7.29). Let $\lambda_1, \dots, \lambda_k$ be the corresponding eigenvalues. Form a $(k \times k)$-matrix P whose i-th column is x_i, and let $D = \mathrm{diag}(\lambda_1, \dots, \lambda_k)$. Then A can be diagonalised as $A = PDP^{-1}$. It follows that for any $n \in \mathbb{N}$, $A^n = PD^nP^{-1}$.

Now, since $x_1, \dots, x_k$ are orthogonal, P is an orthogonal matrix, i.e. $P^\top = P^{-1}$. Hence $A^n = PD^nP^\top$. Note that

$$PD^n = \begin{bmatrix} x_1 \mid \dots \mid x_k \end{bmatrix} \begin{bmatrix} \lambda_1^n & & \\ & \ddots & \\ & & \lambda_k^n \end{bmatrix} = \begin{bmatrix} \lambda_1^n x_1 \mid \dots \mid \lambda_k^n x_k \end{bmatrix},$$

and

$$P^\top v_1 = \begin{bmatrix} x_1 \\ - \\ \vdots \\ - \\ x_k \end{bmatrix} v_1 = \begin{bmatrix} x_1 \cdot v_1 \\ \vdots \\ x_k \cdot v_1 \end{bmatrix},$$

where $\cdot$ denotes the dot product. This means

$$v_{n+1} = A^n v_1 = PD^nP^\top v_1 = \begin{bmatrix} \lambda_1^n x_1 \mid \dots \mid \lambda_k^n x_k \end{bmatrix} \begin{bmatrix} x_1 \cdot v_1 \\ \vdots \\ x_k \cdot v_1 \end{bmatrix} = \sum_{i=1}^{k} (x_i \cdot v_1) \lambda_i^n x_i.$$

We obtain an explicit formula for δ_{n+1}:

$$\delta_{n+1} = [v_n]_p - [v_n]_q = \sum_{i=1}^{k} (x_i \cdot v_1) \lambda_i^n \left([x_i]_p - [x_i]_q\right) = \sum_{i=1}^{k} r_i \lambda_i^n \tag{A.1}$$

where $r_i = (x_i \cdot v_1)\left([x_i]_p - [x_i]_q\right)$. Note that r_i does not depend on n.

Now, it is easy to see that $A = MM^\top$ is *positive semi-definite*, which means its eigenvalues $\lambda_1, \dots, \lambda_k$ are all non-negative. We re-index the sum in (A.1) by grouping together the λ_i which are equal, to get

$$\delta_{n+1} = \sum_{t=1}^{K} R_t \mu_t^n,$$

where $K \leq k$, each R_t is a sum of the r_i (whose corresponding λ_i are equal), and the μ_t are distinct and non-negative. Assume without loss of generality that $\mu_1 > \mu_2 > \cdots > \mu_K \geq 0$. If $R_t = 0$ for all t, then clearly $\operatorname{sign}(\delta_{n+1}) = \operatorname{sign}(0) = 0$ which is constant, so we are done. Otherwise, let T be the minimal t such that $R_t \neq 0$. We may also assume $\mu_T > 0$ (otherwise we necessarily have $\mu_T = 0$, $T = K$ and $\operatorname{sign}(\delta_{n+1}) = \operatorname{sign}(R_T \cdot 0^n)$ which is again constant 0). Observe that

$$\delta_{n+1} = R_T\mu_T^n + \sum_{t=T+1}^{K} R_t\mu_t^n = \mu_T^n \left[R_T + \sum_{t=T+1}^{K} R_t \left(\frac{\mu_t}{\mu_T}\right)^n \right].$$

By our assumption on the ordering of the μ_t, we have $\mu_t < \mu_T$ in the sum. Consequently $|\mu_t/\mu_T| < 1$, and $(\mu_t/\mu_T)^n \to 0$ as $n \to \infty$. This means

$$\lim_{n\to\infty} \left[R_T + \sum_{t=T+1}^{K} R_t \underbrace{\left(\frac{\mu_t}{\mu_T}\right)^n}_{\to 0} \right] = R_T \neq 0.$$

Since this limit is non-zero, there is $m_{pq} \in \mathbb{N}$ such that the sign of term in square brackets is equal to $S = \operatorname{sign} R_T \in \{1, -1\}$ for all $n \geq m_{pq}$. Finally, for such n we have

$$\operatorname{sign} \delta_{n+1} = \operatorname{sign} \left(\underbrace{\mu_T^n}_{>0} \left[R_T + \sum_{t=T+1}^{K} R_t \left(\frac{\mu_t}{\mu_T}\right)^n \right] \right) = \operatorname{sign} \left(R_T + \sum_{t=T+1}^{K} R_t \left(\frac{\mu_t}{\mu_T}\right)^n \right) = S$$

i.e. $\operatorname{sign} \delta_n$ is constant for $n \geq m_{pq} + 1$, as required.

The argument which shows that the difference between claim scores is also eventually constant in sign is almost identical. Write $B = M^\top M$, and observe that $w_{n+1} = B^n w_1$. Since B is also symmetric and positive semi-definite, the proof goes through as above. This completes the proof. □

— B —

Proofs for Chapter 5

B.1 Proof of Theorem 5.6.1

We tackle all three operators at once by first stating a technical result which applies to score-based operators with certain conditions. Since var-based-cond and part-based-cond are score-based (by Proposition 5.3.3), the main result will follow.

Let $\mathcal{P}$ denote the set of partitions of $\mathcal{V}$. Say a score-based operator with prior ranking function r_0 and disagreement function d is *decomposable* via functions $f : \mathcal{P} \to \mathbb{Z}$ and $g : \mathcal{P} \times \mathcal{L}_0 \to \mathbb{N}_0$ if

$$r_0(W) = \sum_{i \in \mathcal{S}} f(\Pi_i^W),$$

$$d(W, \langle i, c, \varphi \rangle) = \begin{cases} g(\Pi_i^W, \varphi), & W, c \models \mathsf{S}_i\varphi \\ \infty, & W, c \not\models \mathsf{S}_i\varphi \end{cases}.$$

In analogy with the **Refinement** property for conditioning operators, say f is *refinement-compatible* if whenever Π refines Π' we have $f(\Pi) \leq f(\Pi')$. Similarly, say g is refinement-compatible if $g(\Pi, \varphi) \leq g(\Pi', \varphi)$ for all $\varphi \in \mathcal{L}_0$ whenever Π refines Π'.

Lemma B.1.1. *Suppose a score-based operator is decomposable via refinement-compatible functions f and g. Then for each $\varphi \in \mathcal{L}_0$ there is a function $d_\varphi : \mathcal{V} \times \mathcal{V} \to \mathbb{R}_{\geq 0}$ such that for any $\mu \in \mathcal{L}_0$ and multiset $\Psi = \{\varphi_1, \ldots, \varphi_n\}$,*

$$\|[B_c^{\sigma_\Psi, \mu}]\| = \operatorname*{argmin}_{v \in \|\mu\|} \sum_{i=1}^{n} \min_{u \in \|\varphi_i\|} d_{\varphi_i}(u, v). \tag{B.1}$$

Proof. For valuations $u, v \in \mathcal{V}$, define a partition $\Pi(u, v)$ by

$$\Pi(u, v) = \{\{u, v\}\} \cup \{\{w\} \mid w \notin \{u, v\}\}.$$

Recall that $\Pi_\perp$ denotes the finest partition $\{\{u\} \mid u \in \mathcal{V}\}$. For $\varphi \in \mathcal{L}_0$, set

$$d_\varphi(u, v) = f(\Pi(u, v)) + g(\Pi(u, v), \varphi) - f(\Pi_\perp) - g(\Pi_\perp, \varphi).$$

Now take some μ and $\Psi = \{\varphi_1, \dots, \varphi_n\}$. Recall that we are in the setting with a single case c, and with sources $\mathcal{S} = \{1, \dots, N, *\}$. Write $[n] = \{1, \dots, n\}$. For any valuation v and $i \in [n]$ such that $v \notin \|\varphi_i\|$, choose some $u_i \in \operatorname{argmin}_{u \in \|\varphi_i\|} d_{\varphi_i}(u, v)$. Then define a world W_v with valuation v and partitions given by

$$\Pi_i^{W_v} = \begin{cases} \Pi_\perp, & i \notin [n] \\ \Pi_\perp, & i \in [n], v \in \|\varphi_i\| \\ \Pi(u_i, v), & i \in [n], v \notin \|\varphi_i\|. \end{cases}$$

Note that u_i depends on Ψ and v in addition to i, but these dependencies are suppressed from the notation.

Claim B.1.1. *$W_v \in \mathcal{X}_{\sigma_{\Psi,\mu}}$ whenever $v \in \|\mu\|$.*

Proof of claim. Since f and g are always finite, it follows from the decomposition property that $W \in \mathcal{X}_\sigma$ iff $W, c \models \mathsf{S}_i\varphi$ for all $\langle i, c, \varphi \rangle \in \sigma$.

Recall that $\sigma_{\Psi,\mu} = (\langle *, c, \mu \rangle, \langle 1, c, \varphi_1 \rangle \dots \langle n, c, \varphi_n \rangle)$. Since $v \in \|\mu\|$, we have $W_v, c \models \mu$ and thus $W_v, c \models \mathsf{S}_*\mu$. Similarly, for $i \in [n]$ we have $W_v, c \models \mathsf{S}_i\varphi_i$ whenever $v \in \|\varphi_i\|$. It remains to consider $i \in [n]$ such that $v \notin \|\varphi_i\|$. By construction we have $\Pi_i^{W_v} = \Pi(u_i, v)$. But $u_i \in \|\varphi_i\|$ by definition, and $\{u_i, v\} \in \Pi(u_i, v)$. Hence $v \in \Pi_i^{W_v}[u_i] \subseteq \Pi_i^{W_v}[\varphi_i]$, i.e. $W_v, c \models \mathsf{S}_i\varphi_i$ as required. $\diamondsuit$

For brevity, write r for the ranking function $r_{\sigma_{\Psi,\mu}}$. Then for $W \in \mathcal{X}_{\sigma_{\Psi,\mu}}$,

$$\begin{aligned} r(W) &= r_0(W) + d(W, \langle *, c, \mu \rangle) + \sum_{i=1}^n d(W, \langle i, c, \varphi_i \rangle) \\ &= \sum_{i \in \mathcal{S}} f(\Pi_i^W) + g(\Pi_\perp, \mu) + \sum_{i=1}^n g(\Pi_i^W, \varphi_i) \\ &= \sum_{i \in \mathcal{S} \setminus ([n] \cup \{*\})} f(\Pi_i^W) + f(\Pi_\perp) + g(\Pi_\perp, \mu) \\ &\qquad + \sum_{i=1}^n \big(f(\Pi_i^W) + g(\Pi_i^W, \varphi_i) \big). \end{aligned} \tag{B.2}$$

Claim B.1.2. *If $W \in \mathcal{X}_{\sigma_{\Psi,\mu}}$ and $v_c^W = v$, then $r(W_v) \le r(W)$.*

Proof of claim. Note that $\mu \in K_c^{\sigma_{\Psi,\mu}}$, so $W, c \models \mu$ and thus $v \in \|\mu\|$. By Claim B.1.1, we may use (B.2). We show that each term in the sums in the expression for $r(W_v)$ is no greater than the corresponding term in $r(W)$.

First, take $i \in \mathcal{S} \setminus ([n] \cup \{*\})$. By construction, $\Pi_i^{W_v} = \Pi_\perp$. Since $\Pi_\perp$ is the finest partition, it certainly refines Π_i^W. By refinement-compatibility of f, we have $f(\Pi_i^{W_v}) \le f(\Pi_i^W)$ as required.

Now take $i \in [n]$. If $v \in \|\varphi_i\|$ then we again have $\Pi_i^{W_v} = \Pi_\perp$, so by refinement-compatibility of f and g we have $f(\Pi_i^{W_v}) + g(\Pi_i^{W_v}, \varphi_i) \leq f(\Pi_i^W) + g(\Pi_i^W, \varphi_i)$.

Suppose instead that $v \notin \|\varphi_i\|$. Since $W \in \mathcal{X}_{\sigma_\Psi,\mu}$, $W, c \models \mathsf{S}_i\varphi_i$. Recalling that the c-valuation of W is v, this means $v \in \Pi_i^W[\varphi_i]$, i.e. there is some $\hat{u} \in \|\varphi_i\|$ such that $v \in \Pi_i^W[\hat{u}]$. It follows that $\Pi(\hat{u}, v)$ refines Π_i^W. By refinement-compatibility,

$$f(\Pi(\hat{u}, v)) + g(\Pi(\hat{u}, v), \varphi_i) \leq f(\Pi_i^W) + g(\Pi_i^W, \varphi_i).$$

By definition, $\Pi_i^{W_v} = \Pi(u_i, v)$. By choice of u_i, we have $d_{\varphi_i}(u_i, v) \leq d_{\varphi_i}(\hat{u}, v)$; considering the definition of d_{φ_i}, this implies

$$f(\Pi(u_i, v)) + g(\Pi(u_i, v), \varphi_i) \leq f(\Pi(\hat{u}, v)) + g(\Pi(\hat{u}, v), \varphi_i).$$

Putting things together, we find $f(\Pi_i^{W_v}) + g(\Pi_i^{W_v}, \varphi_i) \leq f(\Pi_i^W) + g(\Pi_i^W, \varphi_i)$ as required. $\diamondsuit$

Claim B.1.3. *There is a constant C such that for any $v \in \|\mu\|$,*

$$r(W_v) = C + \sum_{i=1}^{n} \min_{u \in \|\varphi_i\|} d_{\varphi_i}(u, v).$$

Proof of claim. Taking off from (B.2), we have

$$\begin{aligned} r(W_v) = &\sum_{i \in \mathcal{S} \setminus ([n] \cup \{*\})} f(\Pi_\perp) + f(\Pi_\perp) + g(\Pi_\perp, \mu) \\ &+ \sum_{i \in [n] : v \in \|\varphi_i\|} (f(\Pi_\perp) + g(\Pi_\perp, \varphi_i)) \\ &+ \sum_{i \in [n] : v \notin \|\varphi_i\|} (f(\Pi(u_i, v)) + g(\Pi(u_i, v), \varphi_i)). \end{aligned}$$

Note that by definition of d_{φ_i},

$$f(\Pi(u, v)) + g(\Pi(u, v), \varphi_i) = d_{\varphi_i}(u, v) + f(\Pi_\perp) + g(\Pi_\perp, \varphi_i).$$

Noting also that $\Pi_\perp = \Pi(v, v)$, we get

$$\begin{aligned} r(W_v) &= (N+1) f(\Pi_\perp) + g(\Pi_\perp, \mu) + \sum_{i \in [n] : v \in \|\varphi_i\|} d_{\varphi_i}(v, v) \\ &\quad + \sum_{i \in [n] : v \notin \|\varphi_i\|} d_{\varphi_i}(u_i, v) + \sum_{i \in [n]} g(\Pi_\perp, \varphi_i) \\ &= C + \sum_{i \in [n] : v \in \|\varphi_i\|} d_{\varphi_i}(v, v) + \sum_{i \in [n] : v \notin \|\varphi_i\|} d_{\varphi_i}(u_i, v), \end{aligned}$$

where $C = (N+1) f(\Pi_\perp) + g(\Pi_\perp, \mu) + \sum_{i=1}^{n} g(\Pi_\perp, \varphi_i)$.

Now, $d_{\varphi_i}(u_i, v) = \min_{u \in \|\varphi_i\|} d(u, v)$ holds by construction of u_i for $i \in [n]$ with $v \notin \|\varphi\|$. For $i \in [n]$ with $v \in \|\varphi_i\|$, it easily observed that $d_{\varphi_i}(v, v) = 0$, and d_{φ_i} is non-negative by refinement-compatibility. Hence $d_{\varphi_i}(v, v) = \min_{u \in \|\varphi_i\|} d_{\varphi_i}(u, v)$ also. Combining the sums above, we find

$$r(W_v) = C + \sum_{i=1}^{n} \min_{u \in \|\varphi_i\|} d_{\varphi_i}(u, v)$$

as required. $\diamondsuit$

We are finally in a position to show (B.1).

$\subseteq$: Take $v \in \|[B_c^{\sigma_\Psi,\mu}]\|$. Then by Lemma 5.4.1 there is $W \in \mathcal{Y}_{\sigma_\Psi,\mu}$ such that $v_c^W = v$. Since $W, c \models \mu$, $v \in \|\mu\|$. Now take any $v' \in \|\mu\|$. Then $r(W) \leq r(W_{v'})$ since W minimises r. But $r(W_v) \leq r(W)$ by Claim B.1.2. Hence $r(W_v) \leq r(W_{v'})$. Applying Claim B.1.3 we find

$$\sum_{i=1}^{n} \min_{u \in \|\varphi_i\|} d_{\varphi_i}(u, v) \leq \sum_{i=1}^{n} \min_{u \in \|\varphi_i\|} d_{\varphi_i}(u, v').$$

Since $v' \in \|\mu\|$ was arbitrary, this shows $v \in \operatorname{argmin}_{v' \in \|\mu\|} \sum_{i=1}^{n} \min_{u \in \|\varphi_i\|} d_{\varphi_i}(u, v')$ as required.

$\supseteq$: Take $v \in \operatorname{argmin}_{v' \in \|\mu\|} \sum_{i=1}^{n} \min_{u \in \|\varphi_i\|} d_{\varphi_i}(u, v')$. We will show $W_v \in \mathcal{Y}_{\sigma_\Psi,\mu}$ and conclude $v \in \|[B_c^{\sigma_\Psi,\mu}]\|$ by Lemma 5.4.1. Indeed, take any $W' \in \mathcal{X}_{\sigma_\Psi,\mu}$. Write $v' = v_c^{W'}$. Then $v' \in \|\mu\|$, so by Claim B.1.3 and our minimising assumption on v, we have $r(W_v) \leq r(W_{v'})$. But using Claim B.1.2 again, $r(W_{v'}) \leq r(W')$. Hence $r(W_v) \leq r(W')$. Since $W' \in \mathcal{X}_{\sigma_\Psi,\mu}$ was arbitrary, we are done. $\square$

The main result follows.

Theorem 5.6.1. $\Delta^{\mathsf{vbc}} \equiv \Delta^{d_H,\Sigma}$, $\Delta^{\mathsf{pbc}} \equiv \Delta^{d_D,\Sigma}$ *and* $\Delta^{\mathsf{exm}} \equiv \Delta^{d_D,\Sigma}$.

Proof. If a score-based operator satisfies the hypotheses of Lemma B.1.1, the associated merging operator Δ has

$$\|\Delta_\mu(\Psi)\| = \operatorname*{argmin}_{v \in \|\mu\|} \sum_{i=1}^{n} \min_{u \in \|\varphi_i\|} d_{\varphi_i}(u, v),$$

for some functions d_φ. This is already close to the form of the model-based merging operator $\Delta^{d,\Sigma}$, except that a different distance function d_{φ_i} is used in each term of the sum. To show that $\mathsf{var\text{-}based\text{-}cond}$ corresponds to $\Delta^{d_H,\Sigma}$ and that $\mathsf{part\text{-}based\text{-}cond}$ and $\mathsf{excess\text{-}min}$ correspond to $\Delta^{d_D,\Sigma}$, it suffices to show that one can replace the minimum of d_{φ_i} with the minimum of d_H or d_D respectively. We take each operator in turn.

1. By Proposition 5.3.3, var-based-cond is score-based. One can easily show it is decomposable via

$$f(\Pi) = -|\{p \in \mathsf{Prop} \mid \Pi[p] = \|p\|\}|,$$
$$g(\Pi, \varphi) = 0.$$

Such f and g are refinement-compatible, and so Lemma B.1.1 applies. Inspecting the proof, we have

$$d_\varphi(u, v) = f(\Pi(u, v)) - f(\Pi_\perp).$$

Clearly this quantity does not depend on φ (since $g \equiv 0$), so we are justified in writing $d(u, v)$ without further qualification. In fact, since $f(\Pi_\perp) = -|\mathsf{Prop}|$, we have

$$\begin{aligned} d(u, v) &= |\mathsf{Prop}| - |\{p \in \mathsf{Prop} \mid \Pi(u, v)[p] = \|p\|\}| \\ &= |\{p \in \mathsf{Prop} \mid \Pi(u, v)[p] \supset \|p\|\}|. \end{aligned}$$

To complete the proof, it suffices to show $d(u, v) = d_H(u, v)$. But this follows immediately upon noticing that $\Pi(u, v)[p] \supseteq \|p\|$ if and only if u and v differ on p.

2. In much the same way as for var-based-cond above, one can show part-based-cond is a decomposable score-based operator via refinement-compatible functions

$$f(\Pi) = -|\Pi|$$
$$g(\Pi, \varphi) = 0.$$

We have

$$d_\varphi(u, v) = f(\Pi(u, v)) - f(\Pi_\perp) = |\mathcal{V}| - |\Pi(u, v)|.$$

It is straightforward to observe that

$$|\Pi(u, v)| = \begin{cases} |\mathcal{V}|, & u = v \\ |\mathcal{V}| - 1, & u \neq v, \end{cases}$$

and thus

$$d_\varphi(u, v) = \begin{cases} 0, & u = v \\ 1, & u \neq v \end{cases}$$

which is equal to the drastic distance $d_D(u, v)$, as required.

3. excess-min satisfies the hypothesis of Lemma B.1.1 by taking

$$f(\Pi) = 0$$
$$g(\Pi, \varphi) = |\Pi[\varphi] \setminus \|\varphi\||.$$

The corresponding distance functions are

$$\begin{aligned} d_\varphi(u,v) &= f(\Pi(u,v)) + g(\Pi(u,v),\varphi) - f(\Pi_\perp) - g(\Pi_\perp,\varphi) \\ &= g(\Pi(u,v),\varphi) \\ &= |\Pi(u,v)[\varphi] \setminus \|\varphi\||. \end{aligned}$$

Unlike in the other two cases, d_φ does in fact depend on φ. To complete the proof, it suffices to show that for any valuation v and any consistent $\varphi \in \mathcal{L}_0$,

$$\min_{u\in\|\varphi\|} d_\varphi(u,v) = \min_{u\in\|\varphi\|} d_D(u,v).$$

First suppose $v \in \|\varphi\|$. Then the minima on both sides are equal, attaining value 0 at $u = v$. If instead $v \notin \|\varphi\|$, then for any $u \in \|\varphi\|$ we have $\Pi(u,v)[\varphi] = \|\varphi\| \cup \{v\}$, so $d_\varphi(u,v) = |\{v\}| = 1$ and the minimum on the left-hand side is 1. But the minimum on the right-hand side is also 1, since $u \in \|\varphi\|$ implies $u \neq v$ and $d_D(u,v) = 1$. This completes the proof.

□