Formal approaches to a definition of agents

submitted to the University of Hertfordshire in partial fulfilment of the requirements

of the degree of PhD

Martin Andreas Biehl

20 January 2017

Contents

Abstract

This thesis is a contribution to the formalisation of the notion of an agent within the class of finite multivariate Markov chains. In accordance with the literature agents are are seen as entities that act, perceive, and are goal-directed. We present a new measure that can be used to identify entities (called *ι-entities*). The intuition behind this is that entities are spatiotemporal patterns for which every part makes every other part more probable. The measure, complete local integration (CLI), is formally investigated within the more general setting of Bayesian networks. It is based on the specific local integration (SLI) which is measured with respect to a partition. CLI is the minimum value of SLI over all partitions. Upper bounds are constructively proven and a possible lower bound is proposed. We also prove a theorem that shows that completely locally integrated spatiotemporal patterns occur as blocks in specific partitions of the global trajectory. Conversely we can identify partitions of global trajectories for which every block is completely locally integrated. These global partitions are the finest partitions that achieve a SLI less or equal to their own SLI. We also establish the transformation behaviour of SLI under permutations of the nodes in the Bayesian network.

We then go on to present three conditions on *general* definitions of entities. These are most prominently not fulfilled by sets of random variables i.e. the perception-action loop, which is often used to model agents, is too restrictive a setting. We instead propose that any general entity definition should in effect specify a subset of the set of all spatiotemporal patterns of a given multivariate Markov chain. Any such definition will then define what we call an entity set. The set of all completely locally integrated spatiotemporal patterns is one example of such a set. Importantly the perception-action loop also naturally induces such an entity set. We then propose formal definitions of actions and perceptions for arbitrary entity sets. We show that these are generalisations of notions defined for the perception-action loop by plugging the *entity-set* of the perception-action loop into our definitions. We also clearly state the properties that general entity-sets have but the perception-action loop entity set does not. This elucidates in what way we are generalising the perception-action loop.

Finally we look at some very simple examples of bivariate Markov chains. We present the disintegration hierarchy, explain it via symmetries, and calculate the ι-entities. Then we apply our definitions of perception and action to these ι-entities.

To my late father and my mother

Acknowledgements

I thank:

Professor Daniel Polani; first and foremost for giving me the chance to pursue this particular line of research; second for great discussions, supervision, criticism, collaboration, and measured encouragement. I also want to thank my second supervisor Professor Chrystopher Nehaniv for support whenever I needed it.

The friends and colleagues at the University of Hertfordshire: Andres Burgos, Christoph Salge, Cornelius Glackin, Nicola Catenacci-Volpi, Lukas Everding, Sander van Dijk, Dari Trendafilov, Martin Greaves, Marcus Scheunemann, Frank Foerster, Antoine Hiolle, Joan Saez.

Professors Takashi Ikegami and Nathaniel Virgo for inviting me as a doctoral fellow of the Japan Society for the Promotion of Science and as a long term visitor to the Earth-Life-Science Institute's Origins Network (EON) respectively. The external interest in my research that I experienced during these two stays in Tokyo were great motivation to continue the path that lead to much of this thesis.

The colleagues and friends at the Earth-Life-Science Institute and Ikegami Lab. Especially: Nicolas Guttenberg, Julien Hubert, Stuart Bartlett, Lana Sinapayen, Olaf Witkowski, Kanjin Yoneda, and Yoichi Mototake.

Professor Florentin Wörgötter for providing a path back into science when it was starting to look unlikely. I also thank him for generously supporting my decision to leave his group when it became clear that our short term goals in research were not aligned closely enough.

The colleagues and friends in Göttingen for discussions and good past times: Frank Hesse, Jan Braun, Christian Tetzlaff, Harm Surkamer, Xiaofeng Xiong, Michael Fauth, Mohammad Aein, Alexey Abramov, Minja Tamosiunaite, Sakyashinga Dasgupta, Tomas Kulvicius, Christoph Kolodziejski, Dennis Goldschmidt, Ahmed Tarek, Christopher Battle, Niko Deuschle, Bernhard Althaner, Clemens Buss, Lukas Geyrhofer, and David Hofmann.

Professors Auke Ijspeert and Karl Svozil for continuing to support my scientific development long after I had been their student. They both also had a profound influence on my thinking.

Basement flat 346 for friendship and hospitality.

Stefanie and Urs Schrade for a great place to write my thesis.

Cornelia and Jan Loewenguth as well as Ulla Biehl and Michael Breuner for supporting my education until late into my thirties.

All the people whose friendship I can rely on even after long phases without any feedback from my side. They are in my heart (even if some of them apparently think that it is grey).

Chapter 1

Introduction

On the most general level this thesis is a contribution to existing research that tries to reconcile a physicalist worldview with the notion of *agents*. The physicalist worldview holds that the laws of physics determine (whether in any way stochastically or not) everything that happens in the universe. The notion of an agent relies fundamentally on the agent's capacity to act. This, however, means that the agent can *make* something happen and that there are not only things happening *to* it (Wilson and Shpall, 2012; McGregor, 2016). It seems that if the agent can make something happen then there must be something that is not determined by the laws of physics. Conversely, if the laws of physics determine everything that happens then the agent did not make it happen. So either the laws of physics do not determine everything or agents do not exist. Let us assume the laws of physics determine everything. We do not know the actual laws of physics but let us also assume that the laws of physics are the same everywhere (in every inertial frame of reference). Then wherever there is a human (the primary example of an agent) and wherever there is no human the laws of physics are the same. These laws of physics do not care about what is happening, they just make it happen. The question remains whether there are agents. Even if humans (or animals, bacteria, plants) cannot really make things happen, our intuition tells us that there is a difference between the volumes of space that contain humans and the volumes of space that do not. The ones that do not contain humans (or other agents) usually are vastly less dangerous for example.

The question is still open what the difference is or even how a difference in danger between volumes of (physically identical) space can arise. In other words it is still an open question (McGregor, 2016) what it is that makes some volumes of space (and their time evolution) *agents*. Specifically human characteristics are not the focus of this research, simple living organisms are sufficient from this point of view.

We want to ascertain that we do not fall prey to our own imagination and give an account of agents that only *seems* compatible with laws of physics. Therefore we choose a completely formal setting. This means we choose a well defined class of "universes" that have laws which are equal basically everywhere. We do not choose the leading theories of physics. Our target here

is the seeming incompatibility between lawfulness and agent containment of a system/universe. There is no need to use complicated systems if we are not sure that the simple ones are not sufficient. We also do not want to assume a priori the existence of notions from physics; most prominently the notion of energy, which gives rise to the notion of work. If it turns out that we need such a concept for agents to exist within a lawful universe then even better. In summary we are firmly in the field of artificial life with its intention to study "life as it could be" (Langton, 1989). More precisely, we study agents as they could be.

The literature (Barandiaran et al., 2009) tells us that agents are entities that act, perceive, and pursue goals. Accordingly, we should try to define each of these notions for a universe governed by (basically everywhere equal) laws of physics. This thesis, building on previous research, proposes formal definitions for *entities*, *action*, and *perception* in such systems. A definition of what it means to pursue goals for entities (that may perceive and act) is not part of this thesis.

As the setting for the formal definitions we choose (possibly driven) multivariate Markov chains. As they include cellular automata, these are suited to model universes with basically everywhere equal laws of physics (Toffoli, 1984). They also include the famous game of life cellular automaton. This is the setting for one of the most complete attempts at a formal definition of agents to date by Beer (2014b). Driven multivariate Markov chains are important because they contain computer implementations of (also continuous) reaction-diffusion systems that exhibit life-like phenomena (Virgo, 2011; Froese et al., 2014; Bartlett and Bullock, 2015, 2016).

Within this setting this thesis splits up into two parts. The first part is the introduction and formal investigation of a newly conceived measure of integration, complete local integration. The second then contains four smaller contributions, a proposition of three requirements for entity definitions, a proposal and motivation of using complete local integration as a definition of entities, a formal definition of actions for arbitrary entities, and a formal definition of perceptions for arbitrary entities.

Apart from Barandiaran et al. (2009) which contains a review of agent definitions we ignore in this thesis all work on agents that is not formal. This means we will not discuss the historical background and philosophical considerations that enable us to even try and formalise agents. We highlight, however, that the formal approaches we are building on are almost all in turn strongly influenced by the work of Maturana and Varela (1980) . This is true for Barandiaran et al. (2009) themselves but equally so for Bertschinger et al. (2006, 2008); Beer (2014a,b). Our work can be seen to a certain extend as a synthesis of these publications.

Recently Beer (2014a,b) has thoroughly investigated the application of the ideas of Maturana and Varela (1980) to the glider in the game of life. In Beer (2014a) he informally introduces criteria for the organisational closure of spatiotemporally extended structures in the game of life (the block, blinker, and glider). From the organisational closure he derives the boundary of the structures and thereby arrives at a definition of entities. It seems to us that this

approach can be further formalised. However, as Beer notes for a formalisation that also accounts for edge cases where closure is temporarily lost, or transformation into other closure regimes further decisions have to be taken. In the end probably for this reason no formal definition of which structures in general constitute entities is given. In Beer (2014b) the glider is then investigated in more detail with respect to its cognitive domain. The cognitive domain contains a notion of perception that we generalise in this thesis and an implicit notion of action which is also similar to the notion of action we propose here. Ignoring goal-directedness this work represents a significant step to a formal definition of agents as it is well defined in a cellular automaton with the same dynamical/physical law at every cell.

A fully formal definition of entities is still missing, however, and the account of perception is also specific to deterministic systems. Furthermore the account of perception (or cognitive domain) may seem peculiar and unrelated to concepts outside the theory of Maturana and Varela (1980).

If we look for formal concepts that can be used as entity definitions outside of Beer's work we find that most candidate notions also have problems discerning entities in edge cases. Such edge cases occur where multiple entities collide, appear or disappear. A common reason for this is that many notions that discern "important" structures/patterns from "less important" structures or more generally just some structures from other structures do so by evaluating structures only spatially. This means they discern among different structures that exist at a single time-step t. Then at the next time-step $t+1$ they again discern between structures at that time-step. The question then remains how to identify which of the structures at t match up with which structures at $t+1$ to form spatiotemporal structures. Often this is unambiguous but when similar structures (like multiple gliders) collide or even overlap then this approach usually fails. An example of such notions are the spatiotemporal filters (Shalizi et al., 2006; Lizier et al., 2008; Flecker et al., 2011) developed for cellular automata. These can highlight gliders, but if two gliders collide they make no claim about the identity of a possibly ensuing glider. Note that these structures were also not conceived for the purpose of detecting entities. The same problem occurs however for the Markov blanket entity underlying the "living organism" of Friston (2013) at least in its current formulation.

An obvious solution to this problem is to directly evaluate spatiotemporal structures for their identity. Then no matching up of time-slices is needed anymore. The existing work identifying such spatiotemporal structures is limited. Balduzzi (2011) detects spatiotemporal coarse-grainings in cellular automata (and multivariate Markov chains in general). This approach may be an alternative to our proposal. It has never been used even for small systems however. Our proposal seems much simpler to express but computationally both are unfeasible for large systems without significant approximations. Another similar work also resulting in a spatiotemporal coarse-graining is the work by Hoel et al. (2013) which identifies causally efficient macrostates. These are however random variables themselves and not spatiotemporal structures like gliders as we will argue. The latter work can be combined with Oizumi et al. (2014), according to the authors, to get spatiotemporal structures more

similar to gliders. In this case after the spatiotemporal coarse-graining of Hoel et al. (2013) the approach of Oizumi et al. (2014) is used to detect spatial structures on top. This looses some flexibility compared to our approach since the coarse-graining does not allow arbitrary combinations of fine-grained spatiotemporal structures anymore. It also does not treat the spatial and temporal dimension on equal footing which is a desirable theoretical property considering the success of relativity theory. These spatiotemporal coarse-grainings have not been specifically proposed as definitions of entities and and do not come with definitions of perception and action. The work by Oizumi et al. (2014) (and predecessors Tononi, 2001; Tononi and Sporns, 2003; Tononi, 2004; Balduzzi and Tononi, 2008) are somewhat related to agents since they try to formally define consciousness. However, for the same reason they identify a single entity (the main complex) in the systems they are used for. These systems are also conceptualised to be applied to neural networks i.e. the "inside" of agents and not to universes to detect the agents. However, investigating the relations of our measure to these will be interesting work for the future.

In summary, currently there is no formally defined and accepted way of identifying entities in multivariate Markov chains. In this thesis we contribute a new measure for this purpose. this measure is called complete local integration (CLI) and we denote the resulting entities as ι-entities. Identifying entities is important for our general research project because it allows to unambiguously and consistently attribute sequences of actions and perceptions over the course of time. This seems to be needed in order to reveal any goal-directed behaviour. This in turn is a defining feature of agents.

The underlying idea of complete local integration is quite simple. We require that every part of an entity makes all other parts of it more probable. Intuitively this can be related to the fact that partial living organisms are extremely rare or at least much rarer than whole living organisms. Not all entities are agents, however, since for example soap-bubbles also have this property[1].

In Chapter 4 we analyse the notion of complete local integration formally. This is done in the general setting of Bayesian networks. These are a generalisation of multivariate Markov chains to cases where notions of time and space are irrelevant or not so simple. This is done since SLI and CLI may be of interest in different contexts as well. First we define the more basic notion of specific local integration with respect to a particular partition. For SLI we constructively prove upper bounds and construct an example of a pattern with strongly negative SLI. These results are of general technical interest and also provide examples. Then we introduce CLI which is the minimum value of SLI with respect to any partition. We then introduce the disintegration hierarchy and the refinement-free disintegration hierarchy. These constructions help reveal the structure of the completely locally integrated patterns and underlie the main formal contribution of this thesis, which is the disintegration theorem (Theorem 22). The disintegration theorem connects the SLI of an entire trajectory (time-evolution) with respect to a partition with the CLI of the blocks of that partition. More precisely for a given trajectory the blocks of the

[1]The author thanks Eric Smith for pointing out this example.

finest partitions among those leading to a particular value of SLI consists only of completely locally integrated blocks. Conversely each completely locally integrated pattern is a block in such a finest partition among those leading to a particular value of SLI. This connection is new. This theorem may lead to further theoretical results and suggests an additional interpretation of completely integrated patterns as independently encoded parts in a code adapted to the specific trajectory (see Section 5.3.5.3). We then go on and investigate the symmetry properties of SLI. We establish its transformation under permutations of the nodes in the Bayesian network in the SLI symmetry theorem and its corollary (Theorems 30 and 31). This can be used to explain the structure of the disintegration hierarchy as we will see in Chapter 6 where we present simple examples. Symmetry properties are also expected to be important for further formal analysis of SLI/CLI. For convenient reference we also show how symmetries spread in multivariate Markov chains, our main application here.

We then come to the second part of this thesis. We have already stated that the notion of perception (part of the cognitive domain) in Beer (2014b) may seem idiosyncratic. However, it turns out to be closely related to the notion of perception that is formalised in the perception-action loop. The perception-action loop is a model of agent-environment interaction that goes back at least to Von Uexküll (1920). Renewed interest possibly started with Beer (1995) and the dynamical systems view of cognition. Later it was formally captured as a Bayesian network by Klyubin et al. (2004) and has been used extensively since then for information theoretic investigations into the interaction of agents and environments (Klyubin et al., 2005; Bertschinger et al., 2006, 2008; Salge et al., 2014; Ay et al., 2012; Zahedi and Ay, 2013).

It is therefore safe to say that the perception-action loop is a powerful tool to investigate such interactions. However, it makes some assumptions that make it unsuitable as a tool for investigating entities. The reason for this is that it models agents as random variables/processes.

We argue in Section 5.3 that a formal notion of entities in multivariate Markov chains should satisfy three criteria. These are *compositionality*, *degree of freedom traversal*, and *counterfactual variation*. It becomes clear in the course of this argument that subsets of the set of random variables in the multivariate Markov chain are not suitable for agent definitions. This includes in particular the perception-action loop since there the agent is just a sequence of random variables.

The three criteria are derived by using what we call the non-preclusion argument. Definitions of entities must allow every phenomenon that is known to be exhibited by any agent (since all agents are entities). For example, if we know that there is a green agent somewhere then an entity definition which says all entities are blue must be wrong. So greenness must not be *precluded* by the entity definition. We argue that, because the glider and other life-like structures in known simulations (Virgo, 2011; Froese et al., 2014; Bartlett and Bullock, 2015; Schmickl et al., 2016) exhibit compositionality, degree of freedom traversal, and counterfactuality both *in value* and *in extent*, entity definitions must not preclude these phenomena. Roughly, compositionality means

that it must be possible that entities have spatial and temporal extension. Degree of freedom traversal (in the game of life for example) means that over time the cells that the entity occupies can change. Counterfactual variation means that entities can be different from one trajectory or time-evolution to another depending on the initial condition for example. Counterfactual variation in value means that there are entities in both trajectories and they occupy exactly the same cells but the occupied cells have different values (e.g. some black ones are white). Counterfactual variation in extent means that the entity in the first trajectory and possibly an entity in the second trajectory occupy different cells. If there is no entity in the second trajectory this is a special kind of counterfactual variation in extent. Apart from ruling out the definition of entities as sets of random variables these three phenomena can help guide future entity definitions. We also believe the non-preclusion argument can be extended to further phenomena such as growth or replication.

The three requirements then convince us that subset of random variables are unsuitable for a general entity definition. We then propose to define entities in general as subsets of the set of spatiotemporal patterns. These are formally defined in Chapter 3 but are basically just subset of the cells with fixed values. Importantly the fixed cells are not limited to one time-step but can spread across arbitrary times. We then call any chosen subset of all spatiotemporal patterns an entity-set. How to arrive at the entity-set is a matter of choice. We propose to use the completely locally integrated spatiotemporal patterns, the ι-entities but our definitions of entity action and entity perception are for arbitrary entity-sets.

These definitions of entity actions and entity perceptions combine ideas from Bertschinger et al. (2008) and Beer (2014b). Let us first come back to the initial problem since we are about to define actions in a lawful system. In a multivariate Markov chain only the transition matrix makes things happen and since all entities are within the chain they cannot possibly make anything happen. Furthermore for each entity in the entity set we are given the full spatiotemporal extension of the entity at once. There is *no choice* for these entities they are completely determined for their entire lifetime. The trick we use to define actions in such a system is to rely on counterfactual entities. That is we use entities that are indistinguishable for the environment. We then say that an entity performs an action at time t if it has a *co-action entity* that cannot be distinguished from the original one by *any* observer in the system. This is ensured if there is a single environment at t that can occur together with both entities. Since the environment is identical nothing in it and therefore no observer can know what the next configuration or time-slice of the entity is. At least if the next time-slices of the co-action entities are actually different. This is another requirement we make of co-action entities. Since the two co-action entities can differ in value or extent at the next time-step we also can differentiate between value and extent actions.

We show that this definition of entity action implies the notion of non-heteronomy due to Bertschinger et al. (2008) in the special case where the entities are the perception-action loop entities. It is possible to show this formally because the perception-loop can be seen as consisting of a special

case of an entity set. This entity set is not composite in space, not degree of freedom traversing, and only counterfactual in value but it is still an entity set (it is also exhaustive which means there is an entity in every trajectory). Due to the generality of the entity set we can therefore treat the perception-action loop as a test case for our definitions. This will also be useful in future research since we can rely on the existing body of work in the perception action-loop and generalise it. The entity set can then serve as a bridge between the perception-action loop and a more general theory of agents in multivariate Markov chains (or even more general Bayesian networks in the future).

We then come to our definition of perception for entity-sets. The basic idea behind entity perception is to capture all influences from the environment on an entity. For the perception-action loop there is a well defined procedure for doing this and we will show this in Section 3.3.6. There we show that we can capture the influence from the environment by a partition of the environment states into states that have the same influence on the agent process. This kind of construction is known in the literature and has been used for example by Balduzzi (2011) in basically the same way. This is also related to the older notion of causal states (Shalizi, 2001). In this thesis we generalise this construction for entity sets. It turns out that the result is also a generalisation of the cognitive domain (more precisely the macroperturbations) defined in Beer (2014b)[2]. So how does this generalisation work? This is formally more involved then we originally envisioned. Again we are forced to deal with the fact that the entities are already defined for their entire lifetimes. So we actually cannot "test" influences on them. Again we rely on other, similar entities to formally capture perception. For a given entity, we take the set of entities that has identical pasts up to some time t. We also make sure that those entities still all exist at $t + 1$. These entities are the co-perception entities. We then classify the environments that can occur with at least one of these entities. These are the co-perception environments. Since the multivariate Markov chains can be stochastic we cannot identify which environment leads to which future (as in Beer, 2014b) we have to do this probabilistically. For this we have to define a probability distribution over the futures of the co-perception entities. In the perception-action loop setting this is straightforward since the futures of the co-perception entities are just the possible values of a random variable[3]. For arbitrary entity sets, the co-perception entities can have undesirable properties. One such property is that they may not exhaustive. This means that the sum of their probabilities does not sum to one as is needed for probability distributions. This can be dealt with in a standard way if the the co-perception entities are mutually exclusive. However, for arbitrary entity sets this is not the case. We then have two options.

- Either we take a subset of the co-perception entities that is mutually exclusive and define the probability distribution over this set. Due to the arbitrary choice of the subset however this leads to a non-unique perceptions; another choice of a subset produce another set of perceptions.

[2]We do not prove this. But we are quite sure.

[3]Due to the special case of the entity-set in the perception-action loop, which consists of all possible combinations of all agent random variables.

- Or we find that the entity set is non-interpenetrating. In that case we can use the whole set of co-perception entities for the definition of the probability distribution. This ensures a uniquely defined set of perceptions. Non-interpenetration is the assumption that two (different) entities with identical pasts cannot occur together. Note that this is not something akin to cell division. Cell division corresponds to a single entity that just becomes two separate *spatial* patterns. Two non-identical entities with the same past that occur together would be more akin to two aligned light beams projected onto a wall unaligning.

In both cases we then arrive at a probability distribution which allows us to classify the co-perception environments. However, this probability distribution is over the entire futures of the co-perception entities. Say there are only two co-perception entities. These may not only be identical up to time t they may be identical up to some arbitrary time $t + r$ in the future. It then seems wrong to interpret the classification of the environments based on the difference in the far future between the two co-perception entities as a perception at time t. To solve this problem we introduce the branching partition. This partitions the co-perception entities according to their next configuration or time-slice. Co-perception entities with equal next configurations are considered as equivalent and part of the same future *branch*. We can then easily derive the probability distribution over these branches by summing over the probabilities in the branch. We call this probability distribution the *branch-morph*.

We then go on to show explicitly that in the special case of the perception-action loop the branch-morph specialises to the standard construction we used to define perceptions in the perception-action loop. We therefore successfully generalise this construction to the case of arbitrary entity sets. In particular these entity sets can be non-exhaustive, degree of freedom traversing, and counterfactual in extent (not only in value). We establish that the branch morph is uniquely defined if the entity-set is non-interpenetrating. This is significant since non-interpenetration then seems like a possible axiom for entity-sets. The branch-morph itself and possibly similar constructions can be used to carry over information theoretic notions from the perception-action loop to entity-sets. This may lead to a definition of goal-directedness. We note here already that the entity set of ι-entities does not satisfy non-interpenetration.

On the technical side we also show how the assumption that all co-perception environments must occur with at least one of the co-perception entities translates to a seemingly weaker requirement in case of the perception-action loop and related cases. This shows conversely that this requirement on the co-perception environments is not stronger than the assumptions made in the perception-action loop case. This is important because we want to generalise the perception-action loop without making extra assumptions.

We then come to the final chapter which presents two extremely simple bivariate Markov chains that have three time-steps. We calculate and visualise the disintegration hierarchy for the first and explain its structure using the SLI symmetry theorems. We also calculate the ι-entities for both chains. We then verify that ι-entities indeed satisfy the three criteria of compositionality, degree

of freedom traversal, and counterfactual variation (in value and extent). We also find some counter intuitive examples of ι-entities.

Then we apply our definitions of action and perception to the calculated ι-entities. We find actions in value and extent. One of the actions we find however seems to question the motivation of our construction this will need further investigation. We also find that ι-entities are generally interpenetrating. We nonetheless construct two branch-morphs based on two different mutually exclusive subsets of the co-perception entities and find that they differ slightly. This is expected.

Finally we discuss the results and give some outlook for future work.

1.1 Original contributions

In summary the original contributions are:

Chapter 4

- Definition of specific local integration (SLI).
- Constructive proof of upper bound of SLI.
- Construction of negative SLI example.
- Definition of complete local integration (CLI).
- Definition of disintegration hierarchy and refinement-free disintegration hierarchy.
- Proof of the disintegration theorem.
- Proof of the SLI symmetry theorems.

Chapter 5

- An argument (via compositionality, degree of freedom traversal, and counterfactual variation) for a spatiotemporal pattern-based definition of entities.
- The abstraction of entity-sets which enables the formal connection to perception-action loop.
- A tentative[4] formal definition of entities as completely locally integrated spatiotemporal patterns.
- A tentative formal definition of action for arbitrary entity-sets.
- A classification of actions into value actions and extent actions.
- A tentative formal definition of perception for arbitrary entity-sets.

[4]For some context on what we mean by "tentative" see Chapter 5.

- An exposition of the role of non-interpenetration of entity-sets in perception. Namely, it makes perception naturally unique.
- The formal exposition of the connection of the action definition to non-heteronomy of Bertschinger et al. (2008) in the perception-action loop.
- The formal exposition of the way the perception definition specialises to the perception-action loop.
- A construction of a conditional probability distribution (the branch-morph, including branching partition) over the futures of entities which allows the definition of perception.
- Proof that the condition on co-perception environments is not stronger than the assumptions about environment states inherent in the perception-action loop.

Chapter 6

- Computation and presentation of disintegration and refinement-free disintegration hierarchies for two simple systems.
- Explanation of the occurrence of multiple disconnected components in the partially ordered disintegration levels via the SLI symmetry theorems.
- Computation and presentation of the completely locally integrated spatiotemporal patterns of two simple systems.
- Examples of ι-entities that exhibit the three phenomena compositionality, degree of freedom traversal, and counterfactual variation that we argued for in Section 5.3.
- Examples of entity actions of ι-entities.
- Example of interpenetrating ι-entities showing that they do not necessarily obey non-interpenetration.
- Example of an entity perception and a branch-morph using a proxy for a co-perception partition.
- Example of an entity action and entity perception of the same ι-entity at the same time-step.
- Discussion of the results on ι-entities as entity sets in the example systems.

Chapter 2

Related work

Here we discuss closely related work in the literature. First we point to the formal origin of the new measures of specific local integration and complete local integration. Then we discuss work that is related to our notion of entities. In cases where the entities are part of conceptions of agents we also discuss perception and action. We have tried to write this chapter without relying too much on our own formalism for accessibility. It might also serve as a further introduction into the field which is why we have left it in front of the technical part of the thesis. Nonetheless, after reading this thesis some arguments will be easier to understand.

2.1 Formally related work

2.1.1 Specific local integration and complete local integration

In Chapter 4 we define specific local integration (SLI). This is a local measure in the sense of the measures of local information dynamics proposed by Lizier (2012). We use the same method of localization presented there only on a different original measure namely multi-information (McGill, 1954; Tononi et al., 1994; Amari, 2001). The method of localising information-theoretic notions like mutual-information and transfer entropy was developed to measure information of specific realisations x, y of random variables X, Y. This is in contrast to the original measures which are averages of the local versions. We argue in Section 5.3.3 that entities should be trajectory dependent. This is equivalent to saying that entities are composed of specific realisations of random variables. Therefore we follow Lizier in using a localised measure.

In contrast to the work by Lizier we are not trying to reveal information storage, transfer and processing to characterise computation within a dynamical system but instead we are trying to find entities and agents within such a system. While the dynamics of information are certainly relevant for agents in dynamical systems we focus here directly on the identification of spatially and temporally composite structures. The measures discussed by Lizier are not designed for the purpose of identifying spatially composite structures (see also Section 2.2.1).

The measure of complete local integration (CLI), which we define in Section 4.3 builds on the notion of SLI. The measure of SLI is defined with respect to a particular partition of a set of random variables. In order to get CLI we evaluate SLI with respect to every possible partition of the set of random variables. We then take the minimum of all the values found in this way to be the value of CLI. A spatiotemporal pattern that has positive CLI value is then defined as an *ι-entity*. The procedure of passing through all partitions has been used for measures similar (and originally equal (Tononi et al., 1994)) to multi-information in Tononi (2001, 2004); Balduzzi and Tononi (2008); Balduzzi (2011); Oizumi et al. (2014). We have adopted it from these publications. Apart from Balduzzi (2011) these publications are part of the integrated information theory approach which we will discuss further below.

It is worth mentioning that our choice of taking the minimum value of SLI found when evaluating all possible partitions of a set of random variables is not without alternatives. Another approach would be to take the (possibly weighted) average of all these values. This has been proposed by Ay (2015) for the non-local multi-information.

In summary the measures of SLI and CLI which we introduce in this thesis are a combination of the idea of localisation of information theoretic measures by Lizier and the origins of integrated information theory by Tononi.

2.2 Work related to our notion of entities

2.2.1 Spatiotemporal filtering and entities

A basic notion in this thesis is that of an entity. At the most basic level the intuition behind this notion is that some spatiotemporal patterns are more important than others. This is also the problem of spatiotemporal filtering. We here discuss work that is similar on this most basic level and then indicate how *ι*-entities essentially differ due to the problem of identity over time.

Defining (and usually finding) more important spatiotemporal patterns or structures (also called coherent structures) has a long history in the theory of cellular automata and distributed dynamical systems. As Shalizi et al. (2006) have argued most of the earlier definitions and methods (Wolfram, 1984; Grassberger, 1984; Hanson and Crutchfield, 1992; Pivato, 2007) require previous knowledge about the patterns being looked for. They are therefor not suitable for a general definition of what entities are. More recent definitions based on information theory (Shalizi et al., 2006; Lizier et al., 2008; Flecker et al., 2011) do not have this limitation anymore. As argued above our method of identifying *ι*-entities is also based on information theoretic notions similar to those used by Lizier et al. (2008). Like the information based definitions in the literature it also requires no knowledge about the system or the patterns that are supposedly interesting. The main difference of our approach is again that it directly results in spatiotemporal patterns and does not go via an intermediate step of evaluating a measure / criterion time-step by time-step. This has certain advantages for our particular purpose.

Applying any one of the definitions (or associated methods) proposed by

(Shalizi et al., 2006; Lizier et al., 2008; Flecker et al., 2011) to the time-evolution (what we call a trajectory) of a cellular automaton assigns each cell (or group of cells) j at each time t a value (usually a real number, but can be discrete as for local statistical complexity in Shalizi et al. (2006)) that measures an important property of the current state of (j, t) [1]. The result is then a "filtered" time evolution of the cells (or groups of cells) in the cellular automaton where each cell j at time t now takes its value of the measured property. These filtered time evolutions then highlight the important spatiotemporally extended structures like gliders and domains. However these methods make no claim about the identity of the revealed patterns. This means that there is no criterion given that tells us which cells and their values at time t_1 and which cells and their values at time t_2 are part of the same entity or object. For isolated gliders this may not seem like a problem but whenever gliders collide it is not clear whether they both loose their identity and become a new thing (or no thing) or whether one of them survived the collision and maybe just changed direction. These questions are not addressed by these publications since the problem of identity over time (or identity of entities in general) is not the focus of these publications. The goal of these publications is to quantify and identify emergent computation and coherent structures and not resolving the identity of entities that may be agents. In order to assign sequences of action and perceptions to entities (or structures) we have to be able to identify them over time. Our approach assigns a measure of integration (CLI) directly to groups of cells that are not only spatially but also temporally extended. We then select the spatiotemporal patterns that have a value above zero as the ι-entities in a given time evolution. If gliders are such ι-entities our approach could make clear whether and which gliders survive collisions.

Note that it could be possible introduce criteria for identity over time via the measured values of the above publications. An example criterion would be to define a threshold and say that all cells whose measured values are above this threshold belong to one entity. However this would often lead to all highlighted structures to be identified as one entity and it is not directly obvious how to define a more detailed entity criterion.

With respect to the criteria for entities we propose in this thesis we find the following

Compositionality Both spatial (e.g. in Shalizi et al. (2006)) and temporal compositionality can occur.

Degree of freedom traversal Degree of freedom traversal can occur. The highlighted spatiotemporal patterns cross from one degree of freedom at one time to another degree of freedom at the next. Just like the gliders they capture.

Counterfactual variation Counterfactual variation can occur. The highlighted spatiotemporal patterns depend on the particular time-evolution (trajectory) of the system.

[1] In the case of local statistic complexity, the value is the causal state not only of the state at (j, t) but of the state of the entire past light-cone. This makes no difference to the following argument however as the result is still just a (discrete) value at (j, t).

Identity Only spatial identity is defined e.g. in Shalizi et al. (2006). Identity over time is not addressed.

Perception, action, goal-directedness There is no intention to define these.

2.2.2 Emergent coarse-graining

The approach most closely related to our own approach and an important inspiration for our work is that of Balduzzi (2011). It proposes a method for coarse-graining the time evolutions (trajectories) of multivariate Markov chains. Using a cellular automaton as an example, the value $x_{j,t}$ of a cell j at each time t is represented by a random variable $X_{j,t}$. This is a common practice in information theoretic/stochastic conceptions of such systems (e.g. Shalizi et al., 2006; Lizier, 2012), which we follow as well. Then, for a given time evolution (x_{out} for Balduzzi and x_V in our formalism), spatiotemporally extended groups of the random variables are combined to form units $\boldsymbol{U}_i$ of the coarse-graining $\mathcal{K}$. The coarse graining $\mathcal{K}$ is formed not only of units but also of ground $\boldsymbol{G}$ and channel $\boldsymbol{C}$. The ground can be related to driving variables (cf. the driven multivariate Markov chain Definition 41) in our case whereas the channel has no analogue in our approach. Ignoring the channel, the coarse-graining is equivalent to a partition of the time evolution like those we investigate in Chapter 4. This means that the units resulting from the coarse-graining method can (by design) be spatiotemporally extended and could correspond to spatiotemporally extended entities that require no additional concept of identity over time. One difference is that the units are also random variables with an associated state space (the coarse-grained alphabet). In our case the ι-entities have a fixed state for all random variables they occupy. They are not random variables themselves. We note that the approach of Balduzzi is then peculiar in the sense that it generates spatiotemporally extended and located coarse-grained random variables that depend on the particular time evolution of a system. This means that our argument against using sets of random variables as agents / entities (see Sections 2.3.1 and 5.3) does not apply to this approach.

The coarse-graining $\mathcal{E}_X(x_{out})$ that best describes the particular time evolution x_{out} for a given system X is also chosen in a way that exhibits some similarities with our approach. First, only *emergent* coarse-grainings are considered. Emergent coarse-grainings satisfy two properties which are too involved to state concisely but which essentially ensure the following:

1. Emergent coarse-grainings are special among the coarse-grainings with equal cardinality. This is makes them similar to refinement-free partitions at a particular disintegration level (see Definition 56).

2. Every unit in these coarse-grainings satisfies a particular condition with respect to its refinements. More precisely, it has more "excess information" than its refinements with respect to the units it is connected to. This is similar to the blocks of the refinement-free partitions at a disintegration level. These are locally integrated with respect to each of their refinements i.e. they have a positive CLI value.

The emergent coarse-grainings are then conceptually somewhat related to the partitions in the refinement-free disintegration hierarchy. One difference is that the units are obtained by looking at how they are connected to other units. In our case we focus only on the internal connection of ι-entities. It would therefore be surprising if the two approaches were measuring the same thing. At the same time it should be noted that excess information as defined by Balduzzi is a partially localised[2] information theoretic measure that considers all possible partitions of the *inputs* of a set of random variables. It is therefore closely related to CLI, which we use. The difference is that CLI partitions the random variables in a group/block/unit directly and not the input variables.

The best coarse-graining $\mathcal{E}_X(x_{out})$ is the one that maximises the "excess information" among all emergent coarse-grainings. A similar requirement could be made in our case by selecting the partition at the lowest level of the refinement-free disintegration hierarchy. We make no such final selection in this thesis but plan to investigate this further in the future.

In summary the approach of Balduzzi has many parallels to our notion of entities (agent properties like actions, perception, or goal-directedness are not treated) and it would be interesting to investigate how the two approaches are related in detail. This will be future work.

With respect to the criteria for entities we propose in this thesis we find the following

Compositionality Both spatial and temporal compositionality can occur. The units are spatiotemporally defined.

Degree of freedom traversal Degree of freedom traversal can occur. Units can cross arbitrarily from one degree of freedom at one time to another degree of freedom at the next.

Counterfactual variation Counterfactual variation can occur. The units depend on the particular time-evolution (trajectory) of the system.

Identity Both spatial and temporal identity are defined in a unified way.

Perception, action, goal-directedness There is no intention to define these.

2.2.3 Integrated information theory

Integrated information theory Tononi (2001); Tononi and Sporns (2003); Tononi (2004); Balduzzi and Tononi (2008); Oizumi et al. (2014) is an attempt to develop a measure of consciousness of physical configurations. Similar to our setting it is defined for the setting of multivariate Markov chains. While the main focus of this theory is consciousness it becomes conceptually related to our work if it is slightly reinterpreted. One of its main goals is to quantify the unity of conscious experiences. In Tononi (2004) the authors also mention that informationally integrated sets form "entities" (also called complexes) that have "ports-in" and "ports-out" to connect to parts that are not within the

[2]Partially localised refers to measure where the averages over some of the random variables in an information theoretic measure are omitted but others are still taken (see Lizier, 2012).

entity. This is very similar to the program of this thesis which is to establish a formal definition of acting and perceiving entities. As far as we know there is no formal definition of these ports-in and ports-out and what constitutes perceptions and actions of them. In its modern formulation (Oizumi et al., 2014) IIT measures the IIT-integration of all spatial patterns $x_{A_t} := (x_{j,t})_{(j,t)\in A_t}$ with $A_t \subset V_t$ [3] at some time-step t. For this all possible partitions of the parent and child nodes in the multivariate Markov chain are evaluated and the minimal value is used to define the IIT-integration of the spatial pattern. This leads to the most integrated patterns which are called complexes. Like in the case of spatiotemporal filtering, no criterion is given as to what patterns at t_1 and what patterns at t_2 belong to the same spatiotemporally extended pattern (or complex). The problem of identity over time is then not solved in this publication. However, the authors refer to Hoel et al. (2013) when mentioning that the spatial patterns should be evaluated over optimal "grains". In Hoel et al. (2013) a method is presented which coarse-grains multivariate Markov chains spatiotemporally. This means that multiple random variables at multiple times are grouped together to form new coarser random variables. Unlike in Balduzzi (2011) these coarse-grainings are not dependent on the particular time-evolution of the chain. They do, however, create also temporally extended structures (random variables) and can therefore be seen to solve the problem of identity over time. If IIT is now used on these coarse-grained random variables we again find the IIT-integrated spatial patterns which are now also temporally extended since the coarse-grained variables are themselves temporally extended on the underlying (not coarse-grained) level. This would lead us to a notion of entity where each entity is a coarse-grained "spatial" pattern that is based on temporally extended underlying patterns. Each such entity / coarse pattern would then correspond to a set of underlying spatiotemporal patterns.

With respect to the criteria for entities we propose in this thesis we find the following

Compositionality Both spatial and temporal compositionality can occur.

Degree of freedom traversal Degree of freedom traversal can occur. The method by Hoel et al. (2013) can create coarse-grained random variables lumping together variables at different times and that belong to different degrees of freedom.

Counterfactual variation A restricted kind of counterfactual variation can occur. IIT evaluates spatial patterns which are values of random variables and therefore change from one time evolution to another. However, we cannot have both degree of freedom traversal and counterfactual variation of the degree of freedom traversal. This means we cannot have full counterfactual variation in extent. More precisely, assume we have two binary degrees of freedom and look at two time steps. Then we have the random variables $\{X_{1,t_1}, X_{2,t_1}, X_{1,t_2}, X_{2,t_2}\}$. Say the coarse-graining selects the two variables X_{1,t_1}, X_{2,t_2} at different times to form a coarse-grained

[3] We write V_t for all random variables in the system at time t.

variable Y then the underlying spatiotemporal patterns exhibit degree of freedom traversal (they switch from the first to the second degree of freedom). These can be identified as IIT integrated if Y is integrated by itself (this is possible). However, now that Y is fixed there can be no entity that does not traverse the degrees of freedom e.g. one occupying only X_{1,t_1} and X_{1,t_2} since these are not together part of a coarse-grained variable and if they are joined via IIT then they must always include all of Y since X_{1,t_1} is part of Y. This means that the coarse-graining restricts the possible counterfactual variation in extent.

Identity Spatial identity is realised by IIT. The coarse graining realises both spatial and temporal identity. The two kinds of identity are therefore not treated in the same way.

Perception, action, goal-directedness There are no formal definitions for these. Parts of the investigated network are sometimes defined as sensor and actuator variables (Albantakis et al., 2014) but in that case the whole network is the "brain" of a animat and not a general universe or biosphere like system.

2.2.4 Kolmogorov complexity of patterns

Recently Zenil et al. (2015) have proposed a method of evaluating spatiotemporal patterns directly (instead of concatenating spatial patterns) by approximating the Kolmogorov complexity. They evaluate 2D patterns (one time and one space dimension) according to the (algorithmic) probability that they are generated by a 2D Turing machine.

The algorithmic probability of one of the patterns is the number of 2D Turing machines that generate the pattern divided by all halting 2D Turing machines. The (Kolmogorov) complexity is then estimated as the self-information (negative logarithm) of this probability. This results in a very general measure for the complexity of patterns. For the purpose of this thesis this approach is too general. We want to explicitly evaluate patterns according to the dynamical laws that generate them i.e. we want to find the spatiotemporal patterns that can be agents within particular multivariate Markov chains. From our point of view some patterns that are agents in one multivariate Markov chain could well be an arbitrary pattern in another chain. If the patterns look the same however the approach of Zenil et al. will ascribe the same value to them independent of the underlying dynamics of the system. It is therefore not applicable to our problem.

With respect to the criteria for entities we propose in this thesis we find the following

Compositionality Both spatial and temporal compositionality can occur.

Degree of freedom traversal Degree of freedom traversal can in principle be evaluated. In the present version however only rectangular patterns are treated this is means there are no degree of freedom traversals.

Counterfactual variation Counterfactual variation can occur. All occurring patterns in a trajectory can be evaluated and these differ in general from trajectory to trajectory. The problem is that all patterns have the same value across all systems/multivariate Markov chains.

Identity Both spatial and temporal identity are defined in a unified way.

Perception, action, goal-directedness There is no intention to define these.

2.3 Work related to our definition of agents (and entities)

2.3.1 Interacting stochastic processes as agents or entities

In the literature it is common to model agents as stochastic (including deterministic) processes interacting with an environment. In its most general formulation this view assumes that at each time t there is a random variable M_t that represents the agent (or its "memory") and a random variable E_t that represents the environment. Interactions can then be modelled via conditional probabilities (see Section 3.3.6). This is also a discretised version of interacting dynamical systems as proposed by Beer (1995) to model agent and environment. Furthermore, this model includes as an important subclass the Markov decision problems and partially observable Markov decision problems Tishby and Polani (2011). Note that in cases of the Markov decision problems the agent memory M_t is often not explicitly modelled but is implicitly assumed to be a part of the system. In the perception-action loop setting various features of agents have been formally investigated. Examples include learning (e.g. reinforcement learning) (Sutton and Barto, 1998), empowerment (Klyubin et al., 2005; Anthony et al., 2009), informational closure (Bertschinger et al., 2006), autonomy (Bertschinger et al., 2008; Seth, 2010), digested information (Salge and Polani, 2011), self-organisation (Ay et al., 2012), thermodynamics of prediction (Still et al., 2012), morphological computation (Zahedi et al., 2010; Zahedi and Ay, 2013), and individuality (Krakauer et al., 2014).

In this thesis we deliberately do not assume that there is a random variable M_t at each time t which corresponds to an agent. Neither do we assume that there is an environment random variable at each time t. We take a multivariate Markov chain whose state at each point in time t is represented by a (finite) set of random variables $\{X_{j,t}\}_{j \in J}$. Whether there exists an agent (or even an entity) at that time is left open. Furthermore, even if there exists an agent at time t it may only exist at time t in one particular time-evolution or trajectory of the system. In another trajectory there might again be no agent at time t or there might be one occupying a different subset of the random variables than in the first case. These situations are not modelled by the perception-action loop framework. They have been ignored or modelled away in order to focus on different aspects of agents. The success of this approach justifies this choice. Since we are interested in a fundamental and general definition

of agents in multivariate Markov chains we cannot follow this choice. We will argue this in more detail in Section 5.3.

Since our definitions also accommodate systems where there is an agent and environment at every time step we will also connect our approach to the perception-action loop after we have defined actions and perception in Section 5.6. In the future we hope that our work contributes to the extension of the work cited above to the more general setting treated here. We take some steps in this direction by generalising perception and action but more detailed investigations are needed to see whether these notions are sufficient.

Among the work cited above we will make use of and are also generally inspired by the fundamental work on the *autonomy* of agents in perception-action loops by Bertschinger et al. (2008). This work has more recently been extended in (Krakauer et al., 2014) where it is proposed as the basis of a method to detect the random variables that represent a biological individual at some time t. Note that also in this newer work, unlike in our case, the individual/agent is assumed to be represented by a set of random variables (not their values) and it is assumed that it is represented by the *same* set of random variables in every time-evolution. Nonetheless, the underlying ideas of Bertschinger et al. autonomy namely non-heteronomy and self-determination both reappear in our conception of agents. The role of self-determination which refers to the influence of the agent's state at one time on its state at a subsequent time is played by the requirement of integration of the ι-entities. We only consider patterns as candidates for agents if their parts are interrelated according to complete local integration. The role of non-heteronomy, which requires that the environment state does not determine the agent's next state is played by our notion of entity action. We relate this notion of action to Bertschinger et al.'s measure of non-heteronomy in Section 5.6.

We also note here that within the formalism of reinforcement learning and in response to the definition of universal intelligence by Legg and Hutter (2007) Orseau and Ring (2012) have argued against the assumption that the agent's random variable M_t (which in this case is seen as the memory/tape of a Turing machine) is guaranteed to exist. The idea there is that in a more realistic setting the environment can also overwrite the agent's memory. They conclude that in the most realistic case there only ever is one memory that the agent's data is embedded in. These arguments for a single system and a blurred boundary between agent and environment then lead to similar conclusions as our arguments for spatiotemporal patterns as entities (that can be agents) in Chapter 5.

Speculating at the end Orseau and Ring propose (also in the setting of cellular automata) to define a utility function which is 1 as long as some chosen "heart" pattern exists and 0 otherwise. The agent is then not further specified but supposed to protect the heart pattern against destructive influence and accordingly regarded the longer it succeeds. The only choice possible is that of the initial condition. We agree that the only choice is the initial condition but a prior choice of a pattern that must be maintained does not seem in accordance with our viewpoint here. Here the kinds of patterns that constitute agents depend on the dynamics of the system / multivariate Markov chain. It

is possible that in some systems having some form of "heart" pattern (a better analogy might be a "gene" pattern) turns out to be just what agents need to persist. This, however, would be a consequence of the dynamics of the chain again and the gene pattern would be the gene pattern under those dynamics and not one that can be chosen externally. The only way we see to make sense of using a "heart" pattern is in a setting where finding the dynamics of the system that preserve it for the maximum amount of time is the goal. This however seems trivial to achieve with dynamics that leave every cell fixed. So for an definition of agents in our setting this approach does not seem to work.

Orseau and Ring then also pose it as open questions in the "one memory" setting what part of the system the agent is and what an agent is (where the boundaries between agent and environment are). This thesis is also an attempt to contribute to the answers to these questions.

With respect to the criteria for entities we propose in this thesis we find the following

Compositionality Both spatial and temporal compositionality can occur. The random variable M_t can be composed out of multiple random variables and it has multiple time-steps.

Degree of freedom traversal Degree of freedom traversal is possible. The random variable M_t can be defined to correspond to a different set of random variables at each time Krakauer et al. (e.g. 2014).

Counterfactual variation Counterfactual variation cannot occur. If the entity is a set of random variables then it is always the same set and only the values change.

Identity Usually in the perception-action loop both spatial and temporal identity are given without any justification. Krakauer et al. (2014) have dropped the spatial assumption and search for the right spatial composition of the individual. Both, spatial and temporal identity could be defined via the coarse-graining method by Hoel et al. (2013). However, no claims have been made that these coarse-grained variables have anything to do with agents.

Perception, action Perception and action are implicitly defined as the interactions between the agent and the environment.

Goal-directedness Goal-directedness is ongoing research. Bertschinger et al. (2008) note that their notion of non-trivial informational closure indicates that the agent has some information about the environment or even models it. This may be related to-goal directedness. Another route is to take cues from inverse reinforcement learning (Ng and Russell, 2000) or work on inferring intentions (e.g. Pantelis et al., 2014).

2.3.2 Autopoiesis and cognition in the game of life

2.3.2.1 Autopoiesis and entities

In Beer (2014a) the author constructs an account of spatiotemporal patterns in the game of life cellular automaton based on the ideas of Maturana and Varela (Varela, 1979; Maturana and Varela, 1980). This can be seen as a definition of entities. Moreover it defines entities as spatiotemporal patterns and therefore the set of all such entities may constitute an entity-set in the sense of our Definition 65. This would make it a direct alternative to our own notion of ι-entities.

The construction of the entities proceeds roughly as follows. First the maps from the Moore neighbourhood to the next state of a cell are classified into five classes of *local processes*. Then these are used to reveal the dynamical structure in the transitions from one time-slice of a spatiotemporal pattern to the next. The used example patterns are the famous block, blinker, and glider and are considered including their temporal extension. Using both the processes and the spatial patterns/values/components (the black and white values of cells are called components) networks characterising the organisation of the spatiotemporally extended patterns are constructed. These can then be investigated for their organisational closure. This is defined to occurs if the same process component relations as before reoccur at a later time. Boundaries of the spatiotemporal patterns are identified by determining the conditions necessary for the reoccurence of the organisation.

Beer mentions that the current version of this method of identifying entities has its limitations. If the closure is perturbed or delayed and then recovered the entity still looses its identity according to this definition. Two possible alternatives are also suggested. The first is to define the *potential for closure* as enough for the ascription of identity. This is questioned as well since a sequence of perturbations can take the entity further and further away from its "defining" organisation and make it hard to still speak of a defining organisation at all. The second alternative is to define that the persistence of any organisational closure indicates identity. It is suggested that this would allow blinkers to transform to gliders.

We note that our definition of ι-entities does not need similar choices to be made since it is not based on the reocurrence of any organisation. As mentioned before, it takes entire spatiotemporal patterns and evaluates their integration. It is then possible that later time-slices of ι-entities have no organisational similarity to earlier ones. This is most similar to the latter proposal where blinkers can transform to gliders. However, in our case not even the blinker or the glider would necessarily need to exhibit a reoccurring organisation explicitly.

It still seems to us that any of the choices proposed by Beer may be used to construct an automatic way to identify autopoietic patterns as a kind of special patterns or entities. By automatic we mean that no knowledge of the structures we are looking for is necessary. Ignoring computational issues again it may be possible to search through all spatiotemporal patterns and look for closures. Once we find a closure we could try to reconstruct the

associated boundaries and thereby obtain complete entities. It is not stated in the paper whether this is possible in principle. If we assume it is then the resulting set of entities is a set of spatiotemporal patterns and therefore an entity-set according to our Definition 65. This means our own definitions of action and perception could be applied to these autopoietic entities. This is not surprising since Beer (2014b) defines a closely related notion of perception himself. This will be discussed next.

2.3.2.2 Cognitive domain and perception

In Beer (2014b) the author constructs (again following Varela, 1979; Maturana and Varela, 1980) the cognitive domain of the glider in the game of life. Our concept of perception can be seen as a generalisation not only of perception in the perception-action loop but also as a generalisation of the cognitive domain in this publication.

To get the cognitive domain Beer (2014b) employs a series of concepts that have analogues in our definition of perception. The glider is defined as an autopoietic entity in the sense of Beer (2014a). This is also a spatiotemporally extended entity x_A in accordance with our definition. First he defines the *microperturbations* $\mathcal{P}$ of the glider. These are the possible states of the boundary around the glider. The set of microperturbations can also be restricted to the nondestructive perturbations. This means those were the glider does not die at the next time-step. The role of the nondestructive microperturbations that preserve the glider identity is played in our case by the co-perception environments $\mathfrak{S}(x_A, t)$ of entity x_A at time t. The set of microperturbations are then classified according to the induced next state of the glider (including its death state if destructive perturbations are allowed[4]). This results in a set of equivalence classes called the *macroperturbations*. In our case these equivalence classes are the *perceptions* of the entity x_A which are the blocks of the *co-perception partition* $\pi^{\mathfrak{S}}(x_A, t)$. In Beer (2014b) the cognitive domain $\mathcal{C}$ is the collection of all macroperturbations of all possible glider states. In our formalism the cognitive domain $\mathcal{C}(x_A)$ of an entity x_A would be the set of all perceptions that occur along the the time-slices of a given entity x_A:

$$\mathcal{C}(x_A) := \{b \in \pi^{\mathfrak{S}}(x_A, t) : A_t, A_{t+1} \neq \varnothing\} \tag{2.1}$$

where the condition $A_t, A_{t+1} \neq \varnothing$ just picks the times where the entity exists at t and $t+1$. If it doesn't exist at $t+1$ then it cannot perceive anything about the environment at t.

The cognitive domain in Beer (2014b) is defined for the autopoietic entities Beer (2014a). Via its macroperturbations it contains a notion of perceptions which is suitable for systems/entity sets that do not contain agents in every trajectory. Recall that this was not the case for perception in the perception action loop. In this thesis we present a generalisation of Beer (2014b)'s macroperturbations to arbitrary entity sets in arbitrary (possibly stochastic) multivariate Markov chains. This reveals the requirement of *non-interpenetration* for entity

[4] We do not use the death state since we don't allow destructive environments.

sets, which allows uniquely defined perception and, accordingly, uniquely defined cognitive domains. Finally we connect the general notion of perception to the perception-action loop setting. This means we also expose a connection between Beer (2014b) and the perception-action loop.

Our work on perception can therefore be regarded as an extension of the cognitive domain notion proposed in Beer (2014b).

2.3.2.3 Summary

With respect to the criteria for entities we propose in this thesis as well as action, perception and goal-directedness we find the following:

Compositionality Both spatial and temporal compositionality can occur (see the glider).

Degree of freedom traversal Degree of freedom traversal is possible (see the glider).

Counterfactual variation Counterfactual variation can occur (see different gliders in different time-evolutions / trajectories).

Identity Temporal identity is defined via the closure condition. Spatial identity is then derived from there. There are some choices left to make. So the notion is not yet unique.

Perception Perception is defined via the macroperturbations. Our notion is a generalisation to arbitrary entities and stochastic settings.

Action There is no explicit definition of action in this work. However, it is mentioned that the sequence of the entity's time-slices i.e. its "behavioural trajectory" would be interpreted as actions by an observer. This is compatible with our notion even if we make an additional explicit requirement. Beer does not require that there must be different possible next time-slices given the *same* environment. Without this requirement the connection of actions to autonomy that we obtain in this thesis is lost. We note that the glider according to our definition can perform an action.

Goal-directedness There is no notion of goal-directedness defined.

2.3.3 Life as we know it

Friston (2013) argues that life is an emergent property of some dynamical systems and that the emergent living organisms are characterised by Markov blankets. Since living organisms are the primary examples of agents we can focus on implicit properties of agents. The Markov blankets define the entities in this publication. This works in the following way. We are given a particle like system [5] with (each two dimensional) position $p^{(i)}$ and velocity

[5] Particles are referred to as subsystems in the original, we deviate from this terminology here. The notation is the original however.

$p'^{(i)}$ for each particle i. The particles also have inner degrees of freedom $\tilde{q}^{(i)}$ but they play no role in the entity definition (they do in persistence etc.). The particles positions and velocities obey some equations of motion that involve the inner degrees of freedom $\tilde{q}$ but apart from this model particles with some friction term in a potential well. The particles only ever interact if they are closer to each other than some threshold (which happens to be equal to 1). From the position of these particles a time dependent adjacency matrix $A(t)$ is derived. The matrix entry $A_{ij}(t)$ is set to 1 if the particles i and j were closer than the threshold within a time-window (of length 256 seconds) preceding t. The matrix $A(t)$ is then used to find the Markov blanket. This is done by constructing the Markov blanket matrix $B(t) := A(t) + A(t)^T + A(t)^T A(t)$ where C^T denotes the transpose of C. At each point in time t the eigenvectors of the Markov blanket matrix $B(t)$ are then calculated. The eigenvector with the largest eigenvalue then contains positive real numbers and indicates in how far the according particle is part of the most interconnected cluster. The particles with the largest $k = 8$ values were then picked to be the internal states i.e. the inside of the entity. So setting $k = 8$ is arbitrary. If we now construct the vector $\chi = (\chi_1, ..., \chi_{128})$ such that $\chi_i = 1$ if i is one of the internal particles (and $\chi_i = 0$ otherwise) the matrix product $A(t)\chi$ will indicate the children, $A(t)^T\chi$ the parents, and $A(t)^T A(t)\chi$ the parents of the children of the internal particles according to the adjacency matrix $A(t)$.

Now let us define:

- $\Lambda(t) \subset \{1, ..., 128\}$ as the set of internal particles at time t,
- $B(t) \subset \{1, ..., 128\}$ as the set of children of the internal particles, the particles of the "active states"[6],
- $S(t) \subset \{1, ..., 128\}$ as the set of parents of the internal particles, the particles of the "sensory states",
- $\Psi(t) \subset \{1, ..., 128\}$ as the rest, the particles of the "external states".[7]

Let us consider position and velocity $\tilde{p}^{(i)} = (p^{(i)}, p'^{(i)})$ and internal degrees of freedom $\tilde{q}^{(i)}$ of each particle i together as one variable $x_i := (\tilde{p}^{(i)}, \tilde{q}^{(i)})$ and let us define for each time t the random variable $X_{i,t}$ to represent the value of x_i at time t. Then an entity is defined at each time t by:

- the internal states $x_{\Lambda(t),t} := (x_{i,t})_{i \in \Lambda(t)}$,
- the active states $x_{B(t),t} := (x_{i,t})_{i \in B(t)}$,
- the sensory states $x_{S(t),t} := (x_{i,t})_{i \in S(t)}$.

Together these form a spatiotemporal pattern $(x_{\Lambda(t),t}, x_{B(t),t}, x_{S(t),t})$ in accordance with our definition.

[6] In the original these are denoted by A but this would be confusing here.

[7] It is not clear to us which set the parents of the children $A(t)^T A(t)\chi$ are supposed to belong to. It is probably either the action states or the sensory states. We ignore them as where they belong to does not affect the reasoning here.

This results in temporally changing entities that can be different from time-evolution to time-evolution. Note that, since $k = 8$ particles are always chosen as the internal states there is always exactly one entity at each time t. If no particles interacted with these $k = 8$ particles in the current time window then there are no active or sensory states. So an entity is an agent if such interactions happen.

Importantly, the problem of identity over time is not addressed. A single Markov blanket is calculated at each time-step this of course leads to a uniquely defined next Markov blanket but why the two share some identity is not argued. In cases where the interconnections between one set of $k = 8$ particles weakens while those of another set of external particles strengthens there can be a discontinuous jump of all internal states at some point. This would still be seen as the time-evolution of the one Markov blanket in the system. Even if multiple Markov blankets were defined there is no proposed method of discerning what would happen in edge cases where Markov blankets collide for example.

In summary this means that the entities of this approach can *vary counterfactually* (they can occupy different random variables from one trajectory to another Definition 71) can traverse degrees of freedom but do not solve the problem of identity over time. They are therefore entities that are more general than the ones in the perception-action loop but less general than some spatiotemporal pattern based entities like the autopoietic entities of Beer (2014a) or our ι-entities.

So much for the entities of this approach. We now discuss perception and action. The condition that interactions have to occur between internal states and other states for there to perception is also contained in our own notion of perception. In our definition whenever there is an influence by the environment on an entity the entity perceives something. This is also the case here. If there is influence by some non-internal states on the internal states these states are defined to be sensory states. This happens since the adjacency matrix will indicate this interaction. Then the set of sensory states becomes non-empty. However, the notion of perception in Friston (2013) includes more than just the existence of sensory states. If the active states are also non-empty then it is argued that the internal states will "appear to solve Bayesian inference about the external states". This is the notion of perception in this publication. It is a more ambitious and higher-level notion of perception than the merely influenced based notion we propose. We note that our notion of perception is not necessarily carried by a set of sensory states (we would say sensory random variables) i.e. in our case there are no random variables that explicitly represent the sensor values/perceptions. Perceptions are a classification of the environment which may or may not be explicitly represented in the system by random variables.

The action states of Friston also have no direct analogue in our concept of action. Our concept of action is deliberately weak and does not even require an influence on the environment. It only requires that the environment does not determine the next state of the entity (non-heteronomy). This condition is also met by Friston's concept of action states as he states that the "flow

of action states does not depend on external states". So if we consider the action states as part of the entity (which is also done by Friston) our notion of actions is exhibited by such entities. Conversely, our entities can exhibit actions without having action-states such that our notion of action is weaker than Friston's.

In contrast to this thesis Friston also argues that the Markov blanket entities if they have sensory and action states exhibit a kind of goal-directedness. He states that "action will appear to conserve the structural and dynamical integrity" of the Markov blanket. If this is true then we have entities, perception, action, and even goal-directedness (the conservation of integrity) and therefore a complete agent definition.

Evaluating these claims is still an open research question.

With respect to the criteria for entities we propose in this thesis as well as action, perception and goal-directedness we find the following:

Compositionality Both spatial and temporal compositionality can occur for the Markov blanket entities.

Degree of freedom traversal Degree of freedom traversal can occur (and does so in the paper). The Markov blanket contains different particles at different times.

Counterfactual variation Counterfactual variation can occur but is restricted. The Markov blankets change from trajectory to trajectory. In the current version, there is always exactly one such entity however. For agents we can say that there is no agent if the internal states of the $k = 8$ internal particles don't interact with any other particles. More than one agent is not defined currently.

Identity Only spatial identity is treated by the Markov blanket. The identity over time is not addressed.

Perception Perception is defined as Bayesian inference about the external states. According to the author it occurs whenever there is a Markov blanket with active states and sensory states.

Action Action is defined via the active states which flow independently of the external states.

Goal-directedness Goal-directedness is emergent since the actions "appear to conserve the structural and dynamical integrity" of the Markov blanket.

Chapter 3

Formal background

In this section the formal background that will be used throughout the thesis is presented. It is assumed the reader is at least vaguely familiar with

- elementary probability theory of discrete random variables, and
- Bayesian networks.

We will nonetheless present many basic definitions for quick reference and to clarify our notation. For the most part we tried to stick to standard notation of probabilities. However the section on symmetries of spatiotemporal patterns (Section 4.5) requires a more elaborate, if more basic, notation than usual. This will be introduced in Section 3.2 and related to more standard notation.

In Section 3.1 we first recall the definitions of partially ordered sets (*posets*) and special posets called *lattices*. Then we look at *partitions* and the *partition lattices* they form. Partitions will be used in multiple ways in this thesis. They are a basic concept in our definitions of specific local interaction in Chapter 4, of ι-entities in Section 5.3.4 and of entity perception in Section 5.5. The partition lattice is also needed for the definition of ι-entities and forms the underlying structure of the disintegration theorem (Theorem 22) which is one of the main contributions of this thesis.

In Section 3.2 we introduce our notation for random variables and probability distributions. We also introduce the notion of *patterns* which is equivalent to the notion of *spatiotemporal patterns*[1]. This simple notion is fundamental to this thesis. Under the name of cylinder sets it is a well known notion that usually does not play a dominant role. In this thesis it is the basis for the formal as well as the conceptual part.

In Section 3.3 we define *Bayesian networks* which are a generalisation of *multivariate Markov chains*. Most of the theorems (including the disintegration theorem and the SLI symmetry theorem) in Chapter 4 hold for Bayesian networks in general and not only for multivariate Markov chains. Multivariate Markov chains are the systems used in the conceptual part of this thesis. There they represent universes or geospheres that may contain agents. A special

[1]Spatiotemporal patterns are just patterns in systems where notions of time and space are defined.

kind of multivariate Markov chains are the driven multivariate Markov chains which are relevant for some applications that involve external influences on a system (e.g. heat baths). In Chapter 4 we prove Theorem 34, which can be used to relate the symmetries of such a driven multivariate Markov chain to the transformation of specific local integration. Section 3.3 also presents the definition of the *perception-action loop* and the *extended perception-action loop*. We will frequently refer to the perception-action loop in the conceptual part since it is a common way to model agents formally. Furthermore, the method of extracting of perceptions from the perception-action loop to get to the extended version is the starting point for our own definition of entity perception in Section 5.5. The proof we present there that the extracted perceptions (or sensor-values) and actions capture all interactions between agent and environment process also supports the interpretation of perception as the total of all influences of the environment on the agent.

This chapter contains no original contributions and only provides the vocabulary and notation for the following. We are not aware of a reference for the proof of Theorem 10 concerning the extraction of perceptions and actions from the perception-action loop. However, we are sure that it is well known among researchers in the field.

3.1 Posets, lattices, and partitions

Here we introduce the terminology of partially ordered sets (short: posets) and state some facts without proofs. We also present the definition of a lattice as a special poset. For a more thorough treatment as well as proofs we refer to Grätzer (2011).

In this thesis, we will mostly use the partial order of "refinement" to relate "set partitions" to each other. These will be defined in Section 3.1.2. All set partitions of a set will turn out to form a lattice. However, we will sometimes look at a subsets of all set partitions and the posets they form. For this reason we start with the more general notion of posets.

3.1.1 Partially ordered sets and lattices

Definition 1 (Partial orders and posets)**.** *A partial order $\preceq$ on a set A is a binary relation that is*

(i) reflexive: $\forall a \in A{:}a \preceq a$,
(ii) antisymmetric: $\forall a, b \in A$: if $a \preceq b$ and $b \preceq a$ then $a = b$, and
(iii) transitive: $\forall a, b, c \in A$: if $a \preceq b$ and $b \preceq c$ then $a \preceq c$.

A together with the partial order $\preceq$ forms a partially ordered set or shorter a poset.

Remarks:

- Technically, a poset is a tuple $\langle A, \preceq \rangle$ where the partial order is explicitly specified. We will usually call A a poset and imply the partial order if it is clear from context.

- A partial order is partial in the sense that there may be elements $a, b \in A$ such that neither $a \preceq b$ nor $b \preceq a$. In the total order defined next this possibility is excluded.

Definition 2. *A total order on a set A is a partial order on A such that for all $a, b \in A$ either $a \preceq b$ or $b \preceq a$.*

A together with a total order $\preceq$ forms a *totally ordered set*.

- The natural numbers together with the usual less or equal relation $\leqslant$ are a totally ordered set.

Definition 3 (Minimal and maximal elements)**.** *A minimal element of a poset A with partial order $\preceq$ is an element $a \in A$ such that if for any $b \in A$ we have $b \preceq a$ then $b = a$. Conversely, a maximal element is an element $c \in A$ such that if for any $b \in A$ we have $c \preceq b$ then $b = c$.*

Remarks:

- A minimal element then has no lesser element within the poset and a maximal element has no greater element within it.
- There can be multiple minimal and maximal elements in a poset.
- Minimal and maximal elements should not be confused with least and greatest elements which we define next.

Definition 4. *Given a poset A an element $a \in A$ is called a least element if for all $b \in A$, $a \preceq b$. We then denote a by $\mathbf{0}$. An element $c \in A$ is called a greatest element if for all $b \in A$, $b \preceq c$. We then denote c by $\mathbf{1}$.*

Remarks:

- The least element is lesser than all elements in the poset and the greatest element is greater than all elements.
- If they exist, least and greatest elements are unique.
- A least (greatest) element is always a minimal (maximal) element but not vice versa.

Definition 5 (Boundedness)**.** *A poset A is bounded if it has both a least and a greatest element.*

Definition 6 (Covering relation)**.** *Given a two elements $a, b \in A$ of a poset A with $a \neq b$ we say b covers a and write $a \preceq\!\!: b$ if there is no $c \in A$ with $a \neq c \neq b$ such that $a \preceq c \preceq b$.*

Remarks:

- If b covers a then b is greater than a and there is no element in between the two.
- If A is finite then knowing all covering relations determines the partial order of A completely (Grätzer, 2011, p.6).

- The join, which is defined below, can be seen as the generalisation of the case where one element covers another to the case where one element "covers" a whole set of elements.
- The covering relation is helpful for the construction of Hasse diagrams of posets. These provide an informative visual impression of partial orders and are introduced next.

Definition 7 (Hasse diagram)**.** *A Hasse diagram is a visualisation of a poset. Given a poset A the Hasse diagram represents the elements of A by dots. The dots representing the elements are arranged in such away that if $a, b \in A$, $a \neq b$, and $a \preceq b$ then the dot representing a is drawn below the dot representing b. An edge is drawn between two elements $a, b \in A$ if $a \prec: b$ i.e. if b covers a. If edges cross in the diagram this does not mean that there is an element of A where they cross and edges never pass through a dot representing an element.*

Remarks:

- No edge is drawn between two elements $a, b \in A$ if $a \preceq b$ but not $a \prec: b$.
- Only drawing edges for the covering relation does not imply a loss of information about the poset since the covering relation determines the partial order completely (see remark to Definition 6).
- For some example Hasse diagrams see Fig. 3.1

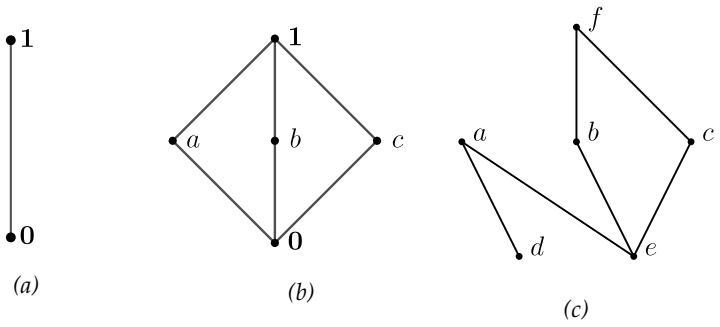


Figure 3.1: Hasse diagrams of three different posets. (a) Two element poset with $\mathbf{0} \preceq \mathbf{1}$. This poset is bounded because it has a zero $\mathbf{0}$ and unit $\mathbf{1}$. (b) A bounded poset with five elements. Note that neither $a \preceq b$ nor $b \preceq a$, so this is a poset but not a totally ordered set. The partial order of this set is $\{\mathbf{0} \preceq a, \mathbf{0} \preceq b, \mathbf{0} \preceq c, \mathbf{0} \preceq \mathbf{1}, a \preceq \mathbf{0}, b \preceq \mathbf{0}, c \preceq \mathbf{0}\}$. The covering relation is $\{\mathbf{0} \prec: a, \mathbf{0} \prec: b, \mathbf{0} \prec: c, a \prec: \mathbf{0}, b \prec: \mathbf{0}, c \prec: \mathbf{0}\}$. Note that the covering relation does not contain a relation between $\mathbf{0}$ and $\mathbf{1}$ but the relation $\mathbf{0} \preceq \mathbf{1}$ (no ":") is a consequence of the covering relation (see remarks to Definitions 6 and 7). (c) A poset without a zero or a unit. Note that d, e are minimal elements and a, f are maximal elements.

Definition 8 (Join and meet)**.** *Given a subset $B \subseteq A$ of elements of a poset A a least upper bound, supremum, or join $\bigvee B$, if it exists, is an element $c \in A$ such that for all $b \in B$ we have $b \preceq c$ and if there exists $a \in A$ such that we also have for all*

$b \in B$ that $b \preceq a$ then $c \preceq a$. Conversely, a greatest lower bound, infimum, or meet $\bigwedge B$, if it exists, is an element $c \in A$ such that for all $b \in B$ we have $c \preceq b$ and if there exists $a \in A$ such that we also have for all $b \in B$ that $a \preceq b$ then $a \preceq c$.

Remarks:

- $\bigvee B$ and $\bigwedge B$ are unique if they exist.
- For a pair of elements $b_1, b_2 \in A$ we also write $b_1 \vee b_2$ for the join and $b_1 \wedge b_2$ for the meet.
- Join and meet are both associative,

$$\pi_1 \wedge (\pi_2 \wedge \pi_3) = \pi_1 \wedge \pi_2 \wedge \pi_3, \tag{3.1}$$
$$\pi_1 \vee (\pi_2 \vee \pi_3) = \pi_1 \vee \pi_2 \vee \pi_3, \tag{3.2}$$

  commutative,

$$\pi_1 \wedge \pi_2 = \pi_2 \wedge \pi_1, \tag{3.3}$$
$$\pi_1 \vee \pi_2 = \pi_2 \vee \pi_1, \tag{3.4}$$

  and idempotent,

$$\pi \wedge \pi = \pi, \tag{3.5}$$
$$\pi \vee \pi = \pi. \tag{3.6}$$

Definition 9 (Lattice)**.** *A poset A is a lattice if for every pair of elements $a, b \in A$ both join and meet exist, i.e. $a \vee b \in A$ and $a \wedge b \in A$.*

Remarks:

- If join and meet exist for every pair they also exist for every finite subset $B \subseteq A$ (Grätzer, 2011, p.9). In this thesis we only encounter finite sets so we have join and meet for every subset, which in general is an additional property of lattices called completeness.
- Every non-empty finite lattice A is a bounded poset with $\mathbf{0} = \bigwedge A$ and $\mathbf{1} = \bigvee A$.
- The posets in Fig. 3.1(a) and Fig. 3.1(b) are lattices.

Definition 10 (Atoms and dual atoms)**.** *Given a bounded poset A an atom is an element $a \in A$ that covers the zero element, i.e. $\mathbf{0} \preceq: a$. A dual atom is an element $b \in A$ that is covered by the unit element, i.e. $b \preceq: \mathbf{1}$.*

3.1.2 Partitions and the partition lattice

This section recalls the definitions of

- set partitions,
- refinement and coarsening of set partitions,

- join and meet operation between partitions,
- the partition lattice.

The following definitions are due to Grätzer (2011, p.359).

Definition 11. *A* (set) partition π *of a set* $\mathcal{X}$ *is a set of non-empty subsets (called* blocks*) of* $\mathcal{X}$ *satisfying*

1. *for all* $x_1, x_2 \in \pi$, *if* $x_1 \neq x_2$, *then* $x_1 \cap x_2 = \varnothing$,
2. $\bigcup_{x \in \pi} = \mathcal{X}$.

We write $\mathfrak{L}(\mathcal{X})$ *for the set of all partitions of* $\mathcal{X}$.

Remark:

- In words, a partition of a set is a set of disjoint non-empty subsets whose union is the whole set.

Definition 12. *If two elements* $x_1, x_2 \in \mathcal{X}$ *belong to the same block of a partition* π *of* $\mathcal{X}$ *write* $x_1 \equiv_\pi x_2$. *Also write* x_1/π *for the block* $\{x_2 \in \mathcal{X} : x_2 \equiv_\pi x_1\}$.

Definition 13 (Refinement and coarsening)**.** *We define the binary relation* $\preccurlyeq$ *between partitions* $\pi, \rho \in \mathfrak{L}(\mathcal{X})$ *as:*

$$\pi \preccurlyeq \rho \text{ if } x_1 \equiv_\pi x_2 \text{ implies } x_1 \equiv_\rho x_2. \tag{3.7}$$

In this case π *is called a* refinement *of* ρ *and* ρ *is called a* coarsening *of* π.

Remark:

- More intuitively, π is a refinement of ρ if all blocks of π can be obtained by further partitioning the blocks of ρ. Conversely ρ is a coarsening of π if all blocks in ρ are unions of blocks in π.
- Refinement and coarsening will be used frequently throughout this thesis.

Theorem 1 (Partition lattice)**.** *Let* $\mathcal{X}$ *be a finite set and* $\mathfrak{L}(\mathcal{X})$ *its set of all partitions.*

(i) Refinement $\preccurlyeq$ *is a partial order of* $\mathfrak{L}(\mathcal{X})$.
(ii) $\mathfrak{L}(X)$ *together with refinement* $\preccurlyeq$ *is a (bounded) lattice. When we write* $\mathfrak{L}(\mathcal{X})$ *in the following we consider it as a lattice with refinement as its partial order.*
(iii) The zero of $\mathfrak{L}(\mathcal{X})$ *is the partition consisting only of singleton sets as blocks,* $\mathbf{0} = \{\{x\}\}_{x \in \mathcal{X}}$.
(iv) The unit of $\mathfrak{L}(\mathcal{X})$ *is the partition consisting of a single block containing all elements,* $\mathbf{1} = \{\mathcal{X}\}$.
(v) The atoms of $\mathfrak{L}(\mathcal{X})$ *are the partitions consisting of a single block containing two elements and apart from this block only of blocks that are singletons.*
(vi) The dual atoms of $\mathfrak{L}(\mathcal{X})$ *are the partitions that consist of two blocks.*
(vii) A partition $\pi \in \mathfrak{L}(\mathcal{X})$ *covers another partition* $\xi \in \mathfrak{L}(X)$, *i.e.* $\xi \preccurlyeq: \pi$ *iff* π *is the result of replacing two blocks of* ξ *by their union.*

Proof. See Grätzer (2011, p.360). □

Remarks:

- The join $\pi_1 \vee \pi_2$ of two partitions $\pi_1, \pi_2 \in \mathfrak{L}(\mathcal{X})$ is the unique coarsening of both π_1 and π_2 which can be obtained by taking the union of the fewest blocks in either of the two partitions. The meet is the partition composed of the blocks obtained by taking the set intersections of all blocks of π_1 and π_2 .
- The Hasse diagrams of the partition lattices of the three and four element set are shown in Figs. 3.2 and 3.3.

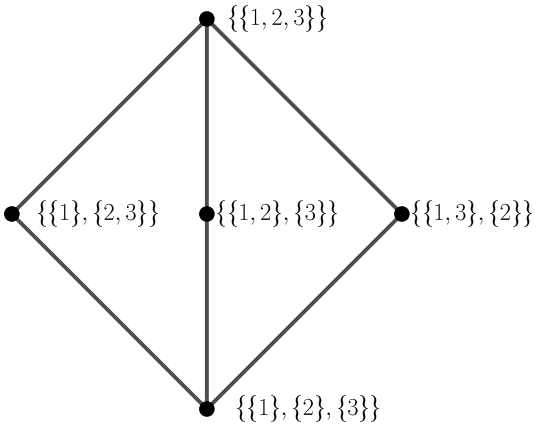


Figure 3.2: Hasse diagrams of the partition lattice of the three element set.

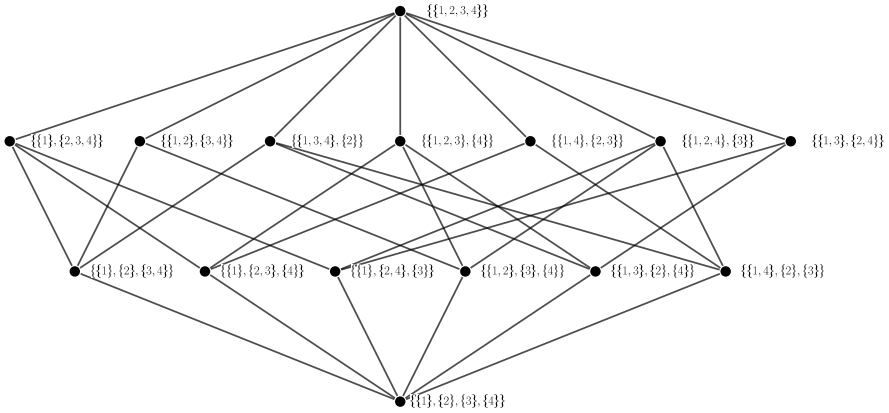


Figure 3.3: Hasse diagrams of the partition lattice of the four element set.

Theorem 2. *The number of partitions of a set $\mathcal{X}$ of cardinality $|\mathcal{X}| = n$ is given by the Bell numbers:*

$$|\mathfrak{L}(\mathcal{X})| = \mathcal{B}_n := \sum_{k=0}^{n} \mathcal{S}(n,k) = \sum_{k=0}^{n} \frac{1}{k!} \sum_{i=0}^{k} (-1)^i \binom{k}{i} (k-i)^n. \tag{3.8}$$

Where $\mathcal{S}(n,k)$ are the Sterling numbers of the second kind which are the number of partitions of a set with n elements into k blocks.

Proof. This is well known. □

3.2 Sets of random variables and patterns

3.2.1 Overview

This section introduces the notion of patterns. In the form of spatiotemporal patterns (which are just patterns in systems where space and time have clear interpretations) this notion forms an important concept in this thesis. The main idea is to capture things/structures that can occur within *single* trajectories of multivariate dynamical systems and stochastic processes. In other words, we would like to capture structures that very from one trajectory of a system to another instead of remaining fixed throughout all trajectories. At the same time these structures should only concern part of a trajectory and not the whole. A reader familiar with probability theory should have no big trouble with the following definitions.

3.2.2 Patterns

Definition 14. *Let $\{X_i\}_{i\in V}$ be set of random variables with index set V and state spaces $\{\mathcal{X}_i\}_{i\in V}$ respectively. Then for $A \subseteq V$ define:*

(i) $X_A := (X_i)_{i\in A}$ *as the random variable composed of the random variables indexed by A,*
(ii) $\mathcal{X}_A := \prod_{i\in A} \mathcal{X}_i$ *as the state space of X_A,*
(iii) $x_A := (x_i)_{i\in A} \in \mathcal{X}_A$ *as a value of X_A,*

Definition 15 (Elementary pattern)**.** *Let $\{X_i\}_{i\in V}$ be set of random variables with index set V and state spaces $\{\mathcal{X}_i\}_{i\in V}$ respectively. An elementary pattern in $\{X_i\}_{i\in V}$ is an assignment*

$$X_j = y$$

of a value $y \in \mathcal{X}_j$ to a single one of the random variables $X_j \in \{X_i\}_{i\in V}$.

We often choose to denote the value $y \in \mathcal{X}_j$ we assign to a random variable X_j by x_j so that it indicates the index j of the random variable X_j it is assigned to. This often allows us to just write x_j for an elementary pattern without danger of ambiguity.

Remark:

- Note that it will later sometimes be necessary to move the values assigned to one random variable to another random variable. This can

result in elementary patterns of the form $X_j = x_k$ where the index k of the assigned value is not identical to the index j of the random variable that it is assigned to. In such cases we will always use the unambiguous full notation $X_j = x_k$ and cannot use just x_k to denote the elementary pattern.

Definition 16 (Pattern)**.** *Let $\{X_i\}_{i \in V}$ be set of random variables with index set V and state spaces $\{\mathcal{X}_i\}_{i \in V}$ respectively. Then a pattern in $\{X_i\}_{i \in V}$ is an assignment*

$$\{X_j = x_j\}_{j \in A}$$

of values $\{x_j\}_{j \in A}$ with $x_j \in \mathcal{X}_j$ for all $j \in A \subseteq V$ to a subset $\{X_j\}_{j \in A}$ of the set of random variables.

An alternative form of specifying a pattern is via $X_A = x_A$ in which case the identical ordering of the tuples $X_A = (X_j)_{j \in A}$ and $(x_j)_{j \in A}$ ensure that the value x_i is assigned to the random variable X_i.

Just like for elementary patterns, if the index set A of a value x_A specifies the index set of the joint random variable that it is assigned to (and in the right order) we use the short notation x_A for the pattern.

Remark:

- Note that by definition a random variable X_i can only occur once and be assigned a single value in a pattern. This is due to the requirement that we can index the elementary patterns that occur in a pattern by a subset $A \subseteq V$. In other words a pattern is just a set of elementary patterns where each random variable occurs on the left hand side in at most one elementary pattern.

- Formally the only difference between the two ways of expressing a pattern is that $\{X_i = x_i\}_{i \in A}$ is a set of elementary patterns and $X_A = x_A \Leftrightarrow (X_i = x_i)_{i \in A}$ is a tuple of elementary patterns. A set is unordered while a tuple is ordered. The expressed assignments are the same.

- This notion of patterns is similar to "patterns" as defined in Ceccherini-Silberstein and Coornaert (2009) and to "cylinders" as defined in Busic et al. (2010). However the notions there are explicitly limited to single time-slices. Our notion of patterns purposely extends to spatiotemporal patterns. These are patterns in specific sets of random variables endowed with notions of time and space like multivariate Markov chains (see Definition 36).

- This notion of patterns is also similar to the notion of propositions. However, propositions are usually more general as they allow any logical combination of the elementary patterns as defined here. For example $X_i = x \; OR \; X_j = y$ is a proposition but not a pattern, since patterns are only defined as lists of elementary patterns not as logical combinations thereof. A set of elementary patterns might be seen as a logical conjunction (AND) of elementary patterns, but a disjunction (OR) is not allowed by the definition.

Definition 17 (Trajectory). *Let $\{X_i\}_{i\in V}$ be set of random variables with index set V and state spaces $\{\mathcal{X}_i\}_{i\in V}$ respectively. A trajectory is a pattern $\{X_j = x_j\}_{j\in V}$ that assigns a value to every random variable in $\{X_i\}_{i\in V}$. We also write x_V for trajectories*

Remark:

- The use of the term "trajectory" might be somewhat surprising at this stage as the set of random variables doesn't have a structure which justifies its use yet. However, the sets of random variables in this thesis will usually correspond to Markov chains which makes trajectory an intuitive choice.
- Trajectories are just particular patterns. However they are special because they determine the whole set of random variables. Since we will need to refer to them frequently it is convenient to give them a particular name.

Definition 18. *Given a set of random variables $\{X_i\}_{i\in V}$ we say pattern x_A occurs in trajectory $\bar{x}_V \in \mathcal{X}_V$ if $\bar{x}_A = x_A$. For formal reasons we also define that all empty patterns are equal i.e. $\bar{x}_\varnothing = x_\varnothing$.*

Remark:

- The definition of $\bar{x}_\varnothing = x_\varnothing$ implies that $x_\varnothing$ occurs in every trajectory.

Definition 19. *Let $\{X_i\}_{i\in V}$ be set of random variables with index set V and state spaces $\{\mathcal{X}_i\}_{i\in V}$ respectively. Given a pattern x_A let $\mathcal{T}(x_A)$ denote the set of trajectories of $\{X_i\}_{i\in V}$ in which x_A occurs i.e.*

$$\mathcal{T}(x_A) := \{\bar{x}_V \in \mathcal{X}_V : \bar{x}_A = x_A\}. \tag{3.9}$$

Remark:

- Note that $\mathcal{T}(x_\varnothing) = \mathcal{X}_V$ because of Definition 18.
- Each pattern thus defines a set of trajectories in which it occurs. However, the converse is not true. As we will see below (Theorem 4) there are subsets of trajectories of a set of random variables that are not captured by any of the patterns in the network. The set of trajectories defined by a pattern is therefore of secondary importance. Primarily patterns are structures that can occur within trajectories.

Theorem 3. *Given a set of random variables $\{X_i\}_{i\in V}$ with index set V and state spaces $\{\mathcal{X}_i\}_{i\in V}$ respectively, the set of all patterns in $\{X_i\}_{i\in V}$ is isomorphic to $\bigcup_{A\subseteq V}\mathcal{X}_A$.*

Proof. Each element $x_B = (x_i)_{i\in B}$ in $\bigcup_{A\subseteq V}\mathcal{X}_A$ defines a pattern $\{X_i = x_i\}_{i\in B}$. And each pattern $\{X_i = x_i\}_{i\in B}$ defines an element $x_B = (x_i)_{i\in B} \in \mathcal{X}_B \subset \bigcup_{A\subseteq V}\mathcal{X}_A$. □

Remark:

- Note $x_\varnothing \in \mathcal{X}_\varnothing \subset \bigcup_{A\subseteq V}\mathcal{X}_A$ is included here since $\varnothing \subset V$. The set $\mathcal{X}_\varnothing$ only contains a single element which is $x_\varnothing$. We could have defined this in an extra definition but it is of no fundamental consequence and so we just note it here.

3.2.3 Patterns and invisible subsets of trajectories

Definition 20 (Anti-pattern)**.** *Given a pattern x_O define its set of anti-patterns $\neg(x_O)$ that have values different from those of x_O on all variables in O:*

$$\neg(x_O) := \{\bar{x}_O \in \mathcal{X}_O : \forall i \in O, \bar{x}_i \neq x_i\}. \tag{3.10}$$

Remark:

- It is important to note that for an element of $\neg(x_O)$ to occur it is not sufficient that x_O does not occur. Only if *every* random variable X_i with $i \in O$ differs from the value x_i specified by x_O does an element of $\neg(x_O)$ necessarily occur. This is why we call $\neg(x_O)$ the anti-pattern of x_O.
- Anti-patterns are useful in the construction of examples of patterns. A generalisation of this anti-patterns will be presented in Definition 48.

Theorem 4. *Given a set of random variables $\{X_i\}_{i\in V}$ where $|V| \geqslant 2$ there are subsets of trajectories $\mathcal{D} \subset \mathcal{X}_V$ such that there is no pattern $x_A \in \bigcup_{C\subseteq V} \mathcal{X}_C$ with $\mathcal{D} = \mathcal{T}(x_A)$.*

Proof. We construct one such subset $\mathcal{D}$ for an arbitrary set of random variables $\{X_i\}_{i\in V}$. Take an arbitrary pattern x_A with $|A| \geqslant 2$ and choose another pattern $\bar{x}_A$ from $\neq (x_A)$. Then let

$$\mathcal{D} := \{\mathcal{T}(x_A) \cup \mathcal{T}(\bar{x}_A)\}. \tag{3.11}$$

To see that there is no pattern $\tilde{x}_B \in \bigcup_{C\subseteq V} \mathcal{X}_C$ with $\mathcal{D} = \mathcal{T}(\tilde{x}_B)$ note that we can write

$$\tilde{x}_C = (\tilde{x}_{C\setminus A}, \tilde{x}_{C\cap A}). \tag{3.12}$$

If $C \cap A \neq \varnothing$ we must have either $\tilde{x}_{C\cap A} = x_A$ or $\tilde{x}_{C\cap A} \neq x_A$. First, let $\tilde{x}_{C\cap A} = x_A$ but then $\mathcal{T}(\bar{x}_A) \nsubseteq \mathcal{T}(\tilde{x}_C)$ so $\mathcal{D} \nsubseteq \mathcal{T}(\tilde{x}_C)$. Next choose $\tilde{x}_{C\cap A} \neq x_A$ but then $\mathcal{T}(x_A) \nsubseteq \mathcal{T}(\tilde{x}_C)$ so also $\mathcal{D} \nsubseteq \mathcal{T}(\tilde{x}_C)$. So we must have $C \cap A = \varnothing$.

Now we show that if $C \cap A = \varnothing$ there are trajectories in $\mathcal{T}(\tilde{x}_C)$ that are not in $\mathcal{D}$. Consider the following trajectory: $\hat{x}_V := (\tilde{x}_C, x_{A_1}, \bar{x}_{A_2}, \check{x}_D)$ where $A_1 \cup A_2 = A$, $A_1 \cap A_2 = \varnothing$, $D = V\setminus(C \cup A)$, and $\check{x}_D \in \mathcal{X}_D$ is arbitrary. We can split up A into A_1 and A_2 like this because $|A| \geqslant 2$ by assumption. Now $\hat{x}_V \in \mathcal{T}(\tilde{x}_C)$ but $\hat{x}_V \neq \mathcal{D}$ because $\hat{x}_A = (x_{A_1}, \bar{x}_{A_2}) \neq x_A$ and $\hat{x}_A \neq \bar{x}_A$ due to our initial choice of $\bar{x}_A \in \neg(x_A)$. □

Remark:

- We explicitly construct a simple example set $\mathcal{D}$ for $V = \{1, 2\}$ and $\{X_i\}_{i\in V} = \{X_1, X_2\}$ the set of random variables. Let $\mathcal{X}_1 = \mathcal{X}_2 = \{0, 1\}$. Then $\mathcal{X}_V = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. Now let $A = V = \{1, 2\}$, choose pattern $x_A = (0, 0)$ and pattern $\bar{x}_A \in \neg(x_A) = (1, 1)$ from its set of anti-patterns. Then let

$$\mathcal{D} := \{\mathcal{T}(x_A) \cup \mathcal{T}(\bar{x}_A)\} = \{(0, 0), (1, 1)\}. \tag{3.13}$$

In this case we can easily list the set of all patterns $\bigcup_{C\subseteq V}\mathcal{X}_C$:

$C \subseteq V$	x_C	$\mathcal{T}(x_C)$
$\varnothing$	$x_\varnothing$	$\mathcal{X}_V$
$\{1\}$	(0)	$\{(0,0),(0,1)\}$
	(1)	$\{(1,0),(1,1)\}$
$\{2\}$	(0)	$\{(0,0),(1,0)\}$
	(1)	$\{(0,1),(1,1)\}$
$\{1,2\}$	$(0,0)$	$\{(0,0)\}$
	$(0,1)$	$\{(0,1)\}$
	$(1,0)$	$\{(1,0)\}$
	$(1,1)$	$\{(1,1)\}$

(3.14)

and verify that $\mathcal{D}$ is not among them. This suggests the first part of the proof above i.e. that $C \cap A = \varnothing$ or else $\mathcal{D} \nsubseteq \mathcal{T}(\tilde{x}_C)$. If there were a further random variable X_3 then any pattern x_3 would contain the trajectory $(x_1, \bar{x}_2, x_3) = (0,1,x_3)$ which is not in $\mathcal{D}$ and corresponds to $\hat{x}_V$ of the proof.

Definition 21 (Visible and invisible subsets)**.** *Given a set of random variables $\{X_i\}_{i\in V}$ a visible subset of $\{X_i\}_{i\in V}$ is a subset of trajectories $\mathcal{D} \subseteq \mathcal{X}_V$ such that there is an pattern x_A in $\{X_i\}_{i\in V}$ with $\mathcal{D} = \mathcal{T}(x_A)$. Subsets $\mathcal{D} \subseteq \mathcal{X}_V$ that are not visible are called invisible subsets of $\{X_i\}_{i\in V}$.*

Remark:

- Visible subsets are completely defined by a pattern whereas this is impossible for invisible subsets. For a given trajectory the question whether a given pattern occurs within this trajectory is well defined via Definition 18. Intuitively this also makes sense since we can just look at the trajectory of a cellular automaton for example to check whether a pattern has occurred. However, for invisible subsets there is no defining pattern, and we cannot inspect a given single trajectory and look for a pattern within it. We can check whether a given single trajectory is an element of the invisible set but this is not the same thing. While this difference might seem to be an inessential subtlety, in this thesis we take this difference seriously.

- The invisible subsets of a set of random variables do not occur within a trajectory in the same way that the patterns do. Let $\mathcal{D}$ be an invisible set. To check whether it occurs in a trajectory $\bar{x}_V$ by the method defined in Definition 18 we cannot directly check whether $\bar{x}_V = \mathcal{D}$ because one is a vector of numbers $(\bar{x}_i)_{i\in V}$ and the other a set of trajectories $\{x_V\}_{x_V\in\mathcal{D}}$. We can check whether $\mathcal{T}(\bar{x}_V) = \mathcal{D}$ (which will always fail by the definition

of invisible sets) but that is not our definition of occurrence. A definition of occurrence that would allow checking all subsets of $\mathcal{X}_V$ would be to require only that $\bar{x}_V \in \mathcal{D}$. Considering our application to Markov chains and dynamical systems, this definition however cannot capture the intuition behind a pattern occurring *in* a single trajectory. This is due to the fact that subsets of trajectories can be constructed from patterns and their anti-patterns together as in the proof of Theorem 4. We would therefore end up in a situation where a subset could "occur in a trajectory" even though it consists of contradictory patterns. Our definitions are specifically designed to talk about occurrences of things (here represented by patterns) within trajectories and not, as the other definition would offer, to talk about trajectories that are contained in subsets. The whole idea behind patterns is to end up being able to talk about things within single trajectories. Maybe it will turn out at some point that patterns are not the right structure for this purpose but in order to evaluate their suitability we are forced to make this distinction between patterns and subsets via the distinction of visible and invisible subsets.

3.2.4 Probabilities of patterns

Definition 22. *Let $\{X_i\}_{i\in V}$ be set of random variables with state spaces $\{\mathcal{X}_i\}_{i\in V}$ respectively and let $A, B \subseteq V$. Also, for all $i \in A \cup B$ let $x_i \in \mathcal{X}_i$. Then:*

(i) The joint probability that the pattern $\{X_i = x_i\}_{i\in A}$ occurs is denoted by:

$$\Pr\left(\{X_i = x_i\}_{i\in A}\right). \tag{3.15}$$

It satisfies the usual conditions:

1.

$$\Pr\left(\{X_i = x_i\}_{i\in A}\right) \in [0,1] \tag{3.16}$$

2.

$$\prod_{j\in A}\sum_{x_j\in\mathcal{X}_j} \Pr\left(\{X_i = x_i\}_{i\in A}\right) = 1. \tag{3.17}$$

(ii) The conditional probability that the pattern $\{X_i = x_i\}_{i\in A}$ occurs given that the pattern $\{X_j = x_j\}_{j\in B}$ occurs is denoted:

$$\Pr\left(\{X_i = x_i\}_{i\in A} \mid \{X_j = x_j\}_{j\in B}\right). \tag{3.18}$$

and defined by:

$$\Pr\left(\{X_i = x_i\}_{i\in A} \mid \{X_j = x_j\}_{j\in B}\right) := \frac{\Pr\left(\{X_i = x_i\}_{i\in A\cup B}\right)}{\Pr\left(\{X_j = x_j\}_{j\in B}\right)}. \tag{3.19}$$

Definition 23 (Probability distribution)**.** *Let $\{X_i\}_{i\in V}$ be set of random variables with index set V and state spaces $\{\mathcal{X}_i\}_{i\in V}$ respectively and let $A \subseteq V$. We define the probability distribution of X_A as the function $p_A : \mathcal{X}_A \to [0,1]$ with*

(i)

$$p_A(x_A) := \Pr(\{X_i = x_i\}_{i\in A}) = \Pr(X_A = x_A) \tag{3.20}$$

Remark:

- The more technically precise term for the probability distribution is a "probability mass function".
- The probability distribution takes the arguments in the order specified by A i.e. the i-th argument is interpreted as the value of the random variable X_i.

Definition 24. *Let $\{X_i\}_{i\in V}$ be set of random variables with index set V and state spaces $\{\mathcal{X}_i\}_{i\in V}$ respectively and let $A, B \subseteq V$ and $A \cap B = \varnothing$. Then we also define:*

$$p_{A,B}(x_A, x_B) := \Pr(X_A = x_A, X_B = x_B) = \Pr(\{X_i = x_i\}_{i\in A}, \{X_i = x_i\}_{i\in B}) \tag{3.21}$$

For convenience we often just write $p_{A\cup B}(x_A, x_B)$ instead of $p_{A,B}(x_A, x_B)$ e.g. if $A \cup B = V$ we write $p_V(x_A, x_B)$ for $p_{A,B}(x_A, x_B)$ this causes no confusion again if the index sets A and B unambiguously indicate the random variables X_A and X_B which they are assigned to.

Remark:

- Note that, technically,

$$p_{A,B}(x_B, x_A) = \Pr(X_A = x_B, X_B = x_A) \neq p_{A,B}(x_A, x_B). \tag{3.22}$$

  In all such cases (where A does not index both the values and the random variables they are assigned to) we will refrain from writing $p_{A\cup B}(x_B, x_A)$. Such cases will arise when we look at symmetries of patterns and their probabilities in Section 4.5. This is the main reason for introducing the more cumbersome full notation of patterns $\{X_i = x_i\}_{i\in A}$ and their probabilities $\Pr(\{X_i = x_i\}_{i\in A})$ in addition to the short notation x_A and the probability distribution p_A.

Definition 25 (Conditional probability distribution)**.** *Let $\{X_i\}_{i\in V}$ be set of random variables with index set V and state spaces $\{\mathcal{X}_i\}_{i\in V}$ respectively and let $A, B \subseteq V$. Then we define:*

$$p_{A|B}(x_A|x_B) := \Pr(X_A = x_A|X_B = x_B) = \Pr(\{X_i = x_i\}_{i\in A} \mid \{X_i = x_i\}_{i\in B}) \tag{3.23}$$

If it is clear from the index set of the values we condition on which random variables they are assigned to, we often just write $p_A(x_A|x_B)$ instead of $p_{A|B}(x_A|x_B)$.

Definition 26 (Morph of a pattern)**.** *Let $\{X_i\}_{i\in V}$ be set of random variables with index set V and state spaces $\{\mathcal{X}_i\}_{i\in V}$ respectively and let x_A be a pattern in $\{X_i\}_{i\in V}$. Then we define the morph denoted by $p_{V\setminus A}(X_{V\setminus A}|x_A)$ of x_A as the probability distribution $p_{V\setminus A|A}(.|x_A) : \mathcal{X}_{V\setminus A} \to [0, 1]$.*

Remark:

- The morph is the probability distribution over the rest of the set of random variables given a pattern x_A.

- This terminology is inspired by Shalizi (2001). In the case where A indicates all past variables in some stochastic process and $V \setminus A$ indicates all future variables, the definition here coincides with the original.

Theorem 5 (Marginalisation). *Let $\{X_i\}_{i \in V}$ be set of random variables with index set V and state spaces $\{\mathcal{X}_i\}_{i \in V}$ respectively and let $A, B \subseteq V$ and $A \cap B = \varnothing$. Then we have:*

$$p_A(x_A) = \sum_{\bar{x}_{V \setminus A}} p_{A, V \setminus A}(x_A, \bar{x}_{V \setminus A}). \tag{3.24}$$

Proof. We do not give a proof here. It follows from the axioms of probability. For a proof we would need to invoke these axioms which is beyond the scope of this thesis. □

Theorem 6 (Chain rule of probability). *Let V be a set of indices for a set of random variables $\{X_i\}_{i \in V}$ with state spaces $\mathcal{X}_i$. Also for all $i \in V$ let $x_i \in \mathcal{X}_i$. Then for any (re-)labelling $i_1, i_2, ..., i_{|V|}$ of the index set V we have:*

$$\Pr\left(\{X_i = x_i\}_{i \in V}\right) = \prod_{j=1}^{|V|} \Pr(\{X_{i_j} = x_{i_j}\} \mid \{X_{i_k} = x_{i_k}\}_{k \in \{j+1, ..., |V|\}}). \tag{3.25}$$

Proof. Follows directly from the definition of the conditional probability, see Definition 22. Just replace the conditional probabilities by their defining fractions and reduce. □

The Kronecker-delta is used in this thesis to represent deterministic conditional distributions.

Definition 27 (Delta). *Let X be a random variable with state space $\mathcal{X}$ then for $x \in \mathcal{X}$ and a subset $C \subset \mathcal{X}$ define*

$$\delta_x(C) := \begin{cases} 1 & \text{if } x \in C, \\ 0 & \text{else.} \end{cases} \tag{3.26}$$

We will abuse this notation if C is a singleton set $C = \{\bar{x}\}$ by writing

$$\delta_x(\bar{x}) := \begin{cases} 1 & \text{if } x \in \{\bar{x}\}, \\ 0 & \text{else.} \end{cases} \tag{3.27}$$

$$= \begin{cases} 1 & \text{if } x = \bar{x}, \\ 0 & \text{else.} \end{cases} \tag{3.28}$$

The second line is a more common definition of the Kronecker-delta.

Remark:

- Let X, Y be two random variables with state spaces $\mathcal{X}, \mathcal{Y}$ and $f : \mathcal{X} \to \mathcal{Y}$ a function such that

$$p(y|x) = \delta_{f(x)}(y), \tag{3.29}$$

then

$$p(y) = \sum_x p_Y(y|x)p_X(x) \tag{3.30}$$
$$= \sum_x \delta_{f(x)}(y)p_X(x) \tag{3.31}$$
$$= \sum_x \delta_x(f^{-1}(y))p_X(x) \tag{3.32}$$
$$= \sum_{x \in f^{-1}(y)} p_X(x) \tag{3.33}$$
$$= p_X(f^{-1}(y)). \tag{3.34}$$

3.3 Bayesian networks

In this section we introduce Bayesian networks and the special cases of it that we will use in this thesis. Our main formal original contributions, the disintegration theorem Theorem 22, and the sli symmetry theorem Theorem 30 in later sections hold for Bayesian networks in general. In the conceptual part of this thesis (Chapter 5) we use multivariate Markov chains (Section 3.3.4) which are a special kind of Bayesian network as systems that can contain agents. A famous example of a deterministic multivariate Markov chain which we also use in Section 5.3 is the game of life cellular automaton. Driven (multivariate) Markov chains (Section 3.3.5) are multivariate Markov chains where the focus is on a subset of the degrees of freedom. Such systems are often used in practice and we therefore include them in our formal considerations. In the conceptual part they lay a lesser role. Finally, in Section 3.3.6 we present the definition of the perception-action loop. This is also a multivariate Bayesian network. The perception-action loop plays a role in this thesis as a reference system that is used to formally represent agents in the literature. Furthermore we present a method to extract perceptions (and actions) from the perception-action loop that capture all influences from the environment on the agent (the actions capture all influences from the agent on the environment). We formally prove that this is the case and in the conceptual part (Section 5.5) we generalize this method of extracting perceptions/influences to "spatiotemporal pattern-based entities".

We therefore present these as special cases of Bayesian networks. For a more thorough treatment of Bayesian networks we refer to Pearl (2000).

3.3.1 Bayesian networks and mechanisms

Definition 28. *A directed acyclic graph $G = (V, E)$ with nodes V and edges E is* factorization compatible *with the joint probabilities the probabilities of a probability distribution $p_V : \mathcal{X}_V \to [0, 1]$ iff the latter can be factorized in the way suggested by G which means:*

$$p_V(x_V) = \prod_{i \in V} p(x_i|x_{\mathrm{pa}(i)}). \tag{3.35}$$

Where $\mathrm{pa}(i)$ *denotes the parents of node* i *according to* G*.*

Remark:

- In general there are multiple directed acyclic graphs that are factorization compatible with the same probability distribution. If we choose any total order for the nodes in V and define a graph by $\mathrm{pa}(i) = \{j \in V : j < i\}$ then Eq. (3.35) becomes Eq. (3.25) which always holds. This means every probability distribution is compatible with all graphs that can be constructed in this way.

Definition 29 (Bayesian network)**.** *A* Bayesian network *is a finite set of random variables* $\{X_i\}_{i\in V}$ *and a directed acyclic graph* $G = (V, E)$ *with nodes indexed by* V *such that the joint probability distribution* $p_V : \mathcal{X}_V \to [0, 1]$ *of* $\{X_i\}_{i\in V}$ *is factorization compatible with* G*. We also refer to the graph set of random variables* $\{X_i\}_{i\in V}$ *as a Bayesian network implying the graph* G*.*

Remark:

- Since $\{X_i\}_{i\in V}$ is finite and G is acyclic there is a set V_0 of nodes without parents.
- We will see specific kinds of Bayesian networks with restricted the graph structures in Sections 3.3.3 to 3.3.5.

Definition 30 (Mechanism)**.** *Given a Bayesian network* $\{X_i\}_{i\in V}$ *with index set* V *for each node with parents i.e. for each node* $i \in V\backslash V_0$ *(with* V_0 *the set of nodes without parents) the mechanism of node* i *or also called the mechanism of random variable* X_i *is the conditional probability (also called a transition kernel)* $p_i : \mathcal{X}_{\mathrm{pa}(i)} \times \mathcal{X}_i \to [0, 1]$ *mapping* $(x_{\mathrm{pa}(i)}, x_i) \mapsto p_i(x_i|x_{\mathrm{pa}(i)})$*. For each* $x_{\mathrm{pa}(i)}$ *the mechanism defines a probability distribution* $p_i(.|x_{\mathrm{pa}(i)}) : \mathcal{X}_i \to [0, 1]$ *satisfying (like any other probability distribution)*

$$\sum_{x_i\in\mathcal{X}_i} p_i(x_i|x_{\mathrm{pa}(i)}) = 1. \tag{3.36}$$

Remark:

- We could define the set of all mechanisms to formally also include the mechanisms of the nodes without parents V_0. However in practice it makes sense to separate the nodes without parents as those that we choose an initial probability distribution over (similar to a boundary condition) which is then turned into a probability distribution p_V over the entire Bayesian network $\{X_i\}_{i\in V}$ via Eq. (3.35). Note that in Eq. (3.35) the nodes in V_0 are not explicit as they are just factors $p_i(x_i|x_{\mathrm{pa}(i)})$ with $\mathrm{pa}(i) = \varnothing$.
- To construct a Bayesian network, take graph $G = (V, E)$ and equip each node $i \in (V\backslash V_0)$ with a mechanism $p_i : \mathcal{X}_{\mathrm{pa}(i)} \times \mathcal{X}_i \to [0, 1]$ and for each node $i \in V_0$ choose a probability distribution $p_i : \mathcal{X}_i \to [0, 1]$. The joint probability distribution is then calculated by the according version of Eq. (3.35):

$$p_V(x_V) = \prod_{i\in V\backslash V_0} p_i(x_i|x_{\mathrm{pa}(i)}) \prod_{j\in V_0} p_j(x_j). \tag{3.37}$$

3.3.2 Deterministic Bayesian networks

Definition 31 (Deterministic mechanism)**.** *A mechanism $p_i : \mathcal{X}_{\mathrm{pa}(i)} \times \mathcal{X}_i \to [0,1]$ is deterministic if there is a function $f_i : \mathcal{X}_{\mathrm{pa}(i)} \to \mathcal{X}_i$ such that*

$$p_i(x_i|x_{\mathrm{pa}(i)}) = \delta_{f_i(x_{\mathrm{pa}(i)})}(x_i) = \begin{cases} 1 & \text{if } x_i = f_i(x_{\mathrm{pa}(i)}), \\ 0 & \text{else.} \end{cases} \tag{3.38}$$

Definition 32 (Deterministic Bayesian network)**.** *A Bayesian network $\{X_i\}_{i\in V}$ is deterministic if all its mechanisms are deterministic.*

Theorem 7. *Given a deterministic Bayesian network $\{X_i\}_{i\in V}$ there exists a function $f_{V\setminus V_0} : \mathcal{X}_{V_0} \to \mathcal{X}_{V\setminus V_0}$ which given a value x_{V_0} of the random variables without parents X_{V_0} returns the value $x_{V\setminus V_0}$ fixing the values of all remaining random variables in the network.*

Proof. According to Eq. (3.35), the definition of conditional probabilities, and using the determinisitic mechanisms we have:

$$p_{V\setminus V_0}(x_{V\setminus V_0}|x_{V_0}) = \prod_{i\in V\setminus V_0} p_i(x_i|x_{\mathrm{pa}(i)}) \tag{3.39}$$

$$= \prod_{i\in V\setminus V_0} \delta_{f_i(x_{\mathrm{pa}(i)})}(x_i). \tag{3.40}$$

For every x_{V_0} the product on the right hand side is a probability distribution and therefore is always greater or equal to zero and maximally one. Also for every x_{V_0} the sum of the probabilities over all $x_{V\setminus V_0} \in X_{V\setminus V_0}$ is equal to one. As a product of zeros and/or ones the right hand side on the second line can only either be zero or one. This means for every x_{V_0} there must be a unique $x_{V\setminus V_0}$ such that the right hand side is equal to one. Define this as the value of the function $f_{V\setminus V_0}(x_{V_0})$. □

Theorem 8 (Pattern probability in a deterministic Bayesian network)**.** *Given a deterministic Bayesian network (Definition 32) and uniform initial distribution $p_{V_0} : \mathcal{X}_{V_0} \to [0,1]$ the probability of the occurrence of an pattern x_A is:*

$$p_A(x_A) = \frac{N(x_A)}{|\mathcal{X}_{V_0}|} \tag{3.41}$$

where $N(x_A)$ is the number of trajectories $\bar{x}_V$ in which x_A occurs.

Proof. Recall that in a deterministic Bayesian network we have a function $f_{V\setminus V_0} : \mathcal{X}_{V_0} \to \mathcal{X}_{V\setminus V_0}$ (see Theorem 7) which maps a given value of x_{V_0} to the value of the rest of the network $x_{V\setminus V_0}$. We calculate $p_A(x_A)$ for an arbitrary subset $A \subset V$. To make this more readable let $A \cap V_0 = A_0$, $A\setminus V_0 = A_r$,

$B := V \backslash A$, $B \cap V_0 = B_0$, and $B \backslash V_0 = B_r$. Then

$$p_A(x_A) = \sum_{\bar{x}_B} p_V(x_A, \bar{x}_B) \tag{3.42}$$

$$= \sum_{\bar{x}_{B_0}, \bar{x}_{B_r}} p_V(x_{A_r}, \bar{x}_{B_r} | x_{A_0}, \bar{x}_{B_0}) p_{V_0}(x_{A_0}, \bar{x}_{B_0}) \tag{3.43}$$

$$= \sum_{\bar{x}_{B_0}, \bar{x}_{B_r}} \delta_{f_{V \backslash V_0}(x_{A_0}, \bar{x}_{B_0})}(x_{A_r}, \bar{x}_{B_r}) p_{V_0}(x_{A_0}, \bar{x}_{B_0}) \tag{3.44}$$

$$= \sum_{\bar{x}_{B_r}} \sum_{\{\bar{x}_{B_0} : (x_{A_0}, \bar{x}_{B_0}) \in f^{-1}_{V \backslash V_0}(x_{A_r}, \bar{x}_{B_r})\}} p_{V_0}(x_{A_0}, \bar{x}_{B_0}) \tag{3.45}$$

$$= \frac{1}{|\mathcal{X}_{V_0}|} \sum_{\bar{x}_{B_r}} |\{\bar{x}_{B_0} \in \mathcal{X}_{B_0} : (x_{A_0}, \bar{x}_{B_0}) \in f^{-1}_{V \backslash V_0}(x_{A_r}, \bar{x}_{B_r})\}| \tag{3.46}$$

$$= \frac{1}{|\mathcal{X}_{V_0}|} N(x_A) \tag{3.47}$$

In the second to last line we used the uniformity of the initial distribution p_{V_0}. The second sum in the second to last line counts all initial conditions that are compatible with x_{A_0} and lead to the occurrence of x_{A_r} together with some $\bar{x}_{B_r}$. The first one then sums over all such $\bar{x}_{B_r}$ to get all initial conditions that are compatible with x_{A_0} and lead to the occurrence of x_{A_r}. Together these are all initial conditions compatible with x_A. In a deterministic system the number of initial conditions that lead to the occurrence of an pattern x_A is equal to the number of trajectories $N(x_A)$ since every different initial condition will produce a single, unique trajectory. □

Remark:

- Due to the finiteness of the network, deterministic mechanisms, and chosen uniform initial distribution the minimum possible non-zero probability for an pattern x_A is $1/|\mathcal{X}_{V_0}|$. This happens for any pattern that only occurs in a single trajectory. Furthermore the probability of any pattern is a multiple of $1/|\mathcal{X}_{V_0}|$.

3.3.3 Univariate Markov chain

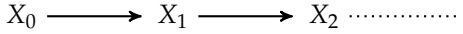


Figure 3.4: First time steps of the Bayesian network representing a univariate Markov chain $\{X_t\}_{t \in T}$.

Markov chains are commonly used to model processes that have no memory of past states. The Bayesian network[2] in Fig. 3.4 shows three initial time steps of a univariate Markov chain. This is a discrete time stochastic process.

[2]We abuse the terminology here by also referring to the directed acyclic graph G associated to a Bayesian network as a "Bayesian network". Context always resolves this ambiguity however.

This means that the index set V is isomorphic to a contiguous subset of the integers. We assume that V is also finite such that we can also assume (without further loss of generality) $V := T = \{0, ..., n-1\}$ with $n \in \mathbb{N}^+$. Then the defining feature is that for all $t \in T$ the random variable with index $t+1$ only depends on the random variable indexed by t. We also assume that the state spaces $\mathcal{X}_t$ of the random variables X_t are all equal. From the graph in Fig. 3.4 we can read:

$$p_T(x_T) = \prod_{t \in T} p_t(x_t | x_{\mathrm{pa}(t)}), \tag{3.48}$$

$$= \prod_{t=1}^{n-1} p_t(x_t | x_{t-1}) p_0(x_0), \tag{3.49}$$

where p_0 is the initial distribution. From Eq. (3.48) we can see the usual Markov chain condition:

$$p_{t+1}(x_{t+1} | x_{\leq t}) = p_{t+1}(x_{t+1} | x_t), \tag{3.50}$$

where $x_{\leq t} = (x_i)_{i \leqslant t}$ is the entire history up to and including t. Starting from a Bayesian network we can then define the Markov chain as follows.

Definition 33. *Let $T = \{0, 1, ..., n-1\}$, $n \in \mathbb{N}^+$. Then a univariate Markov chain is a Bayesian network with random variables $\{X_t\}_{t \in T}$ and graph $G = (T, E)$ such that*

- $\forall t_1, t_2 \in T : \mathcal{X}_{t_1} = \mathcal{X}_{t_2}$,
- $\mathrm{pa}(0) = \varnothing$,
- $\forall t \in T \backslash \{0\} : \mathrm{pa}(t) = \{t-1\}$.

We can also write the right hand side of Eq. (3.50) as a Matrix called a Markov matrix.

Definition 34. *Given a Markov chain $\{X_i\}_{i \in V}$ the Markov matrix P_t at time step t is the matrix with entries:*

$$(P_t)_{x_t, x_{t-1}} := p_t(x_t | x_{t-1}). \tag{3.51}$$

Correspondingly we can write the probability distribution over any random variable X_t as a column vector $p_t = (p_t(x_t))_{x_t \in \mathcal{X}_t}$. Then we can propagate thes probability distributions forward in time using the Markov matrix:

$$p_{t+1} = P_{t+1} p_t. \tag{3.52}$$

An especially simple case of Markov chains are time-homogenous Markov chains.

Definition 35. *A time-homogenous Markov chain is a Markov chain with index set T such that for all $t_1, t_2 \in V$*

$$P_{t_1} = P_{t_2}. \tag{3.53}$$

So the dynamics of time-homogenous Markov chains do not change over time.

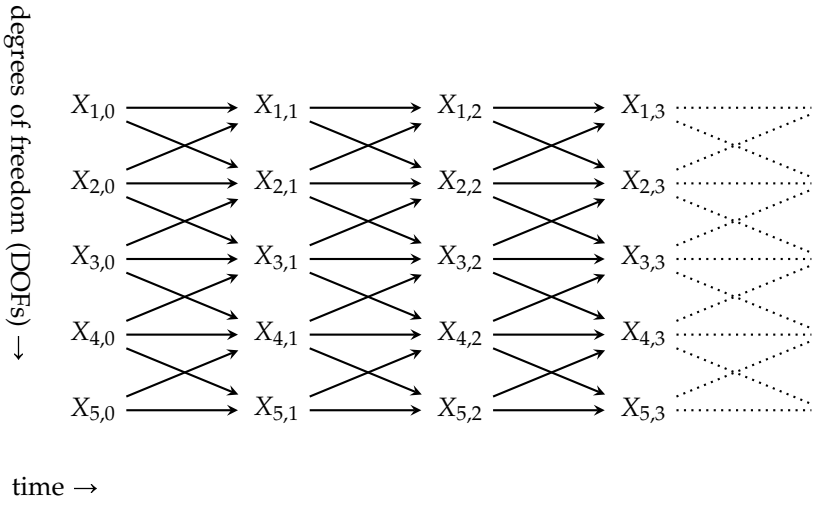


Figure 3.5: First time steps of the Bayesian network representing a multivariate Markov chain $\{X_t\}_{t\in T}$. The shown edges are just an example, any two nodes within the same or subsequent columns can be connected.

3.3.4 Multivariate Markov chain

A multivariate Markov chain is just a Markov chain where the random variable X_t at each point in time is replaced by a whole set of random variables. We also define their time-slices here which will be used throughout this thesis. We then show that time-slices obey the Markov property (this is not surprising and only included for technical reference). Finally we define the notions of space- and time-homogeneity, and instantaneous interactions.

Definition 36 (Multivariate Markov chain)**.** *Let $T = \{0, ..., n\}$ with $n \in \mathbb{N}^+$, $J = \{1, ..., d\}$ with $d \in \mathbb{N}^+$ and $V = \{(j,t) \in J \times T\}$. Then a multivariate Markov chain is a Bayesian network with random variables $\{X_i\}_{i\in V}$ and graph $G = (V, E)$ such that*

- $\mathrm{pa}(j,0) \subseteq \{(k,0) \in V : k \in J \backslash j\}$,
- $\mathrm{pa}(j,t) \subseteq \{(k,t) \in V : k \in J \backslash j\} \cup \{(k,t-1) : k \in J\}$.

We call j the spatial index and t the temporal index.

Remark:

- In essence we get a set of random variables $\{X_{j,t}\}_{j\in J}$ at each point in time t which only depends either on variables at the same point in time or the previous point in time $t-1$ (see Fig. 3.5 for the graph structure). This ensures that the joint random variable $(X_{j,t})_{j\in J}$ composed of all random variables at time t only depends on the joint random variable $(X_{j,t-1})_{j\in J}$ composed of all random variables at time $t-1$ which ensure the Markov property of the joint random vairables. Since these joint

random variables occur repeatedly throughout this thesis we introduce a specific terminology and notation for them.

- Dependencies among random variables $\{X_{j,t}\}_{j\in J}$ at the same point in time t are explicitly allowed in our definition of multivariate Markov chains as they do not break the Markov property as we will see in Theorem 9. Note also that the graph of the entire Bayesian network of the multivariate Markov chain is still directed and acyclic so that we have no cycles among random variables at the same point in time either.

Definition 37 (Time-slices)**.** *Let $V = J \times T$ be an index set composed of a spatial index J and a temporal index T then:*

(i) The time-slice V_t of V at time t is the set of indices

$$V_t := \{(j,t) \in V : j \in J\}. \tag{3.54}$$

(ii) Similarly, for any subset $A \subseteq V$ the time-slice A_t of A at time t is the set of indices

$$A_t := \{(j,t) \in A : j \in J\}. \tag{3.55}$$

(iii) Given a multivariate Markov chain $\{X_j\}_{j\in V}$ with index set V and a subset $A \subset V$ of indices the time-slice X_{A_t} of X_A at time t is the joint random variable indicated by the time-slice A_t of A

$$X_{A_t} := (X_j)_{j\in A_t}. \tag{3.56}$$

Theorem 9. *The time-slices X_{V_t} of multivariate Markov chains $\{X_i\}_{i\in V}$ of Definition 36 satisfy a the Markov property and therefore form a Markov chain. Formally, for all $t \in T$:*

$$p_{V_{t+1}}(x_{V_{t+1}}|x_{V_{\leq t}}) = p_{V_{t+1}}(x_{V_{t+1}}|x_{V_t}). \tag{3.57}$$

Proof. According to Definition 36 the parents of each node are composed of two subsets

$$\begin{aligned} \mathrm{pa}(j,t+1) &= \{(k,t+1) \in V : k \in J\backslash j\} \cup \{(k,t) : k \in J\} && (3.58)\\ &= (\mathrm{pa}(j,t+1) \cap V_{t+1}) \cup (\mathrm{pa}(j,t+1) \cap V_t). && (3.59) \end{aligned}$$

This means that using $i = (j, t+1)$ we can write

$$\begin{aligned} p_{V_{t+1}}(x_{V_{t+1}}|x_{V_{\leq t}}) &= \prod_{i\in V_{t+1}} p_i(x_i|x_{\mathrm{pa}(i)}) && (3.60)\\ &= \prod_{i\in V_{t+1}} p_i(x_i|x_{(\mathrm{pa}(i)\cap V_{t+1})\cup(\mathrm{pa}(i)\cap V_t)}) && (3.61)\\ &= p_{V_{t+1}}(x_{V_{t+1}}|x_{(\bigcup_{i\in V_{t+1}}\mathrm{pa}(i))\backslash V_{t+1}}) && (3.62)\\ &= p_{V_{t+1}}(x_{V_{t+1}}|x_{V_t}). && (3.63) \end{aligned}$$

□

Definition 38. *Given a multivariate Markov chain $\{X_i\}_{i\in V}$ the Markov matrix P_t at time step t is the matrix with entries:*

$$(P_t)_{x_{V_t},x_{V_{t-1}}} := p_{V_t}(x_{V_t}|x_{V_{t-1}}). \tag{3.64}$$

Definition 39 (Space- and time-homogeneity)**.** *A multivariate Markov chain with random variables $\{X_i\}_{i\in V}$ and index set $V = J \times T$ is*

(i) time-homogenous if for all $t_1, t_2 \in T$

1. $\mathcal{X}_{V_{t_1}} = \mathcal{X}_{V_{t_2}}$,
2. *for all $x_{V_{t_1}}, x_{V_{t_2}} \in \mathcal{X}_{V_{t_1}}$ we have*

$$p_{V_{t_1+1}}(x_{V_{t_1+1}}|x_{V_{t_1}}) = p_{V_{t_2+1}}(x_{V_{t_2+1}}|x_{V_{t_2}}), \tag{3.65}$$

or in terms of the Markov matrices:

$$P_{t_1} = P_{t_2}; \tag{3.66}$$

(ii) space-homogenous if for all $j_1, j_2 \in J$ and all $t \in T$

1. $\mathcal{X}_{j_1,t} = \mathcal{X}_{j_2,t}$,
2. $\mathcal{X}_{\mathrm{pa}(j_1,t)} = \mathcal{X}_{\mathrm{pa}(j_2,t)}$,
3. *for all $x \in \mathcal{X}_{j_1,t}$, and $y \in \mathcal{X}_{\mathrm{pa}(j_1,t)}$ we have*

$$p_{j_1,t}(x|y) = p_{j_2,t}(x|y). \tag{3.67}$$

Remark:

- So space-homogeneity means that all mechanisms associated to the random variables within a given time-slice are the same.

Definition 40. *An instantaneous interaction is an edge $(X_{j,t}, X_{k,t}) \in E$ between random variables within the same time-slice.*

Remarks:

- An example of a multivariate Markov chain without instantaneous interactions is the perception-action loop of Section 3.3.6.
- Examples of space- and time-homogenous, deterministic, multivariate Markov chains without instantaneous interactions include the elementary cellular automata (Wolfram, 1983) as well as the Game of Life cellular automaton (Conway, 1970).

3.3.5 Driven processes

A driven process can model systems under the influence of changing external factors or control parameters. An example for such a system is the geosphere which is driven by influence from cosmic and solar radiation. At the same time earth also emits radiation into the cosmos. Acccordingly, in general we allow interaction in both directions with the driving process $\{Y_t\}_{t\in T}$. To define

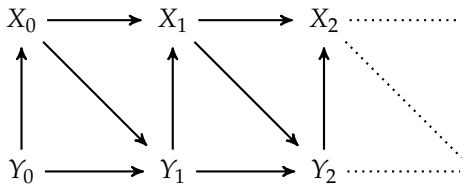


Figure 3.6: First time steps of the Bayesian network representing a process $\{Y_t\}_{t\in T}$ driving a process $\{X_t\}_{t\in T}$.

the driven process $\{X_t\}_{t\in T}$ we define a Bayesian network with two interacting processes and rename one of them $\{Y_t\}_{t\in T}$ in order to simplify discussions. We assume that both processes have constant state spaces. For a visualization of the graph of the according Bayesian network see Fig. 3.6.

Definition 41 (Driven process)**.** *Let $T = \{0, ..., n\}$ with $n \in \mathbb{N}^+$ and $V = \{(j,t) \in \{1,2\} \times T\}$. Consider the Bayesian network with:*

- $\forall (j,t_1),(j,t_2) \in V : \mathcal{X}_{j,t_1} = \mathcal{X}_{j,t_2}$
- $\mathrm{pa}(1,0) := \varnothing$,
- $\mathrm{pa}(2,0) := \{(1,0)\}$,
- $\forall t \in T\backslash\{0\} : \mathrm{pa}(1,t)) = \{(1,t-1),(2,t-1)\}$,
- $\forall t \in T\backslash\{0\} : \mathrm{pa}(2,t)) = \{(1,t),(2,t-1)\}$.

Rename:

- $\{X_{1,t}\}_{t\in T} =: \{Y_t\}_{t\in T}$,
- $\{X_{2,t}\}_{t\in T} =: \{X_t\}_{t\in T}$.

Then $\{Y_t\}_{t\in T}$ is called the driving process and $\{X_t\}_{t\in T}$ the driven process.

Remark:

- We note that $(X_t, Y_t)_{t\in T}$ is a bivariate Markov chain.
- We choose the driving process to interact instantaneously with the driven process as a convention. The main aspect of a driven process is that it highlights the possibility to pay particular attention to the driven process's dynamics and ignore those of the driving process.

Definition 42 (Driven multivariate Markov chain)**.** *Let $T = \{0, ..., n\}$ with $n \in \mathbb{N}^+$, $J = A \cup B$ with $A \cap B = \varnothing$, and $V = \{(j,t) \in J \times T\}$. Consider the Bayesian network with:*

- $\forall (j,t_1),(j,t_2) \in V : \mathcal{X}_{j,t_1} = \mathcal{X}_{j,t_2}$
- *for $j \in B$,* $\mathrm{pa}(j,0) := \varnothing$,

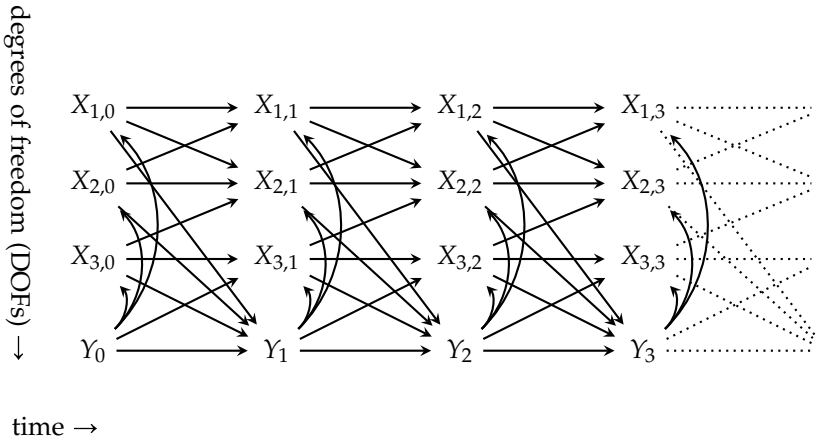


Figure 3.7: First time steps of the Bayesian network representing a multivariate process $\{X_t\}_{t\in T}$ driven by a process $\{Y_t\}_{t\in T}$. Note that the process $\{Y_t\}_{t\in T}$ can also be multivariate, but this would further clutter the graph. Also note that not all edges depicted here must be present. Here, each random variable in each time-slice of the driven process is influenced by the driving process and influences it.

- *for* $j \in A$, $\mathrm{pa}(j,0) \subset \{(k,0) : k \in B\}$,
- $\forall t \in T\backslash\{0\}, j \in B : \mathrm{pa}(j,t)) \subset \{(k,t-1) : k \in A \cup B\}$,
- $\forall t \in T\backslash\{0\}, j \in A : \mathrm{pa}(j,t)) \subset \{(k,t-1) : k \in A\} \cup \{(k,t) : k \in B\}$.

Then $\{X_{B_t}\}_{t\in T}$ is called the driving chain and $\{X_{A_t}\}_{t\in T}$ the (multivariate) driven Markov chain.

Remark:

- We note that $(X_{V_t})_{t\in T}$ is a multivariate Markov chain.
- If $B = \varnothing$ then $\{X_i\}_{i\in V} = \{X_{A_t}\}_{t\in T}$ is a multivariate Markov chain.
- A trajectory of the driven Markov chain is a pattern $x_{A,T} \in \mathcal{X}_{A,T}$ where $(A,T) := A \times T$.
- The multivariate driven Markov chain models a situation where the focus is on the multivariate process $\{X_{A_t}\}_{t\in T}$ even if it is influenced by other (usually very simple) processes contained in $\{X_{B_t}\}_{t\in T}$.
- From the definition we also can write:

$$p_{V_{t+1}}(x_{V_{t+1}}|x_{V_t}) = p_{A_{t+1}}(x_{A_{t+1}}|x_{B_{t+1}}, x_{A_t})p_{B_{t+1}}(x_{B_{t+1}}|x_{A_t}, x_{B_t}). \quad (3.68)$$

- See Fig. 3.7 for an example of a driven multivariate Markov chain.

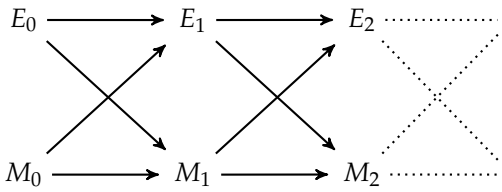


Figure 3.8: First timesteps of the Bayesian network of the perception-action loop. The processes represent environment $\{E_t\}_{t\in T}$, and agent memory $\{M_t\}_{t\in T}$.

3.3.6 Perception-action loop

Here we formally introduce the Bayesian network of the perception-action loop. We first introduce a simple version and then show how to extract implied actions and perceptions without altering the global probability distribution over agent and environment processes. The perception-action loop has been employed to define informational closure (Bertschinger et al., 2006), autonomy (Bertschinger et al., 2008), and morphological computation (Zahedi and Ay, 2013) of agents. In Section 5.6 we discuss the relaiton of our concept of agents to the agent concept that is implicit in the perception-action loop.

Conceptually perception-action loops go back at least to Von Uexküll (1920). Recent formalizations of the perception-action loop (also perception-action cycle, sensorimotor loop) due to Beer (1995) using dynamical systems and Klyubin et al. (2004) using Bayesian networks. Since then it has also been employed, sometimes with minor alterations, by Bertschinger et al. (2006, 2008); Zahedi et al. (2010); Salge and Polani (2011); Ay and Zahedi (2014).

For a rough intuition think of the *perception-action* as modelling the interactions between an "agent" and its "environment" that occur over time. Roughly speaking, at each time-step the agent influences the environment via actions and the environment influences the agent via the latter's perceptions (sensors). It is important to note that in the formal models *all* interactions between agent and environment are captured by the sensor values and actions (Beer, 1995).

Let us first consider a simple perception-action loop consisting of agent and environment only. We assume here that there are no instantaneous interactions between agent and environment. This is particularly suitable for the situation where we relate our conception of agents to that underlying the perception-action loop in Section 5.6. In the perception-action loop the agent memory is represented by a sequence of random variables $\{M_t\}_{t\in T}$, the environment state by $\{E_t\}_{t\in T}$. At each time-step the agent memory M_t is influenced by the last memory state m_{t-1} and the last environment state M_{t-1}. Conversely the environment state E_t is influenced by E_{t-1} and M_{t-1}. See Fig. 3.8 for the Bayesian network. Next, we define the perception-action loop formally.

Definition 43 (Perception-action loop)**.** *Let $T = \{0, ..., n\}$ with $n \in \mathbb{N}^+$ and $V = \{(j,t) \in \{1,2\} \times T\}$. Consider the Bayesian network $\{X_i\}_{i\in V}$ with:*

- $\mathrm{pa}(1,0) := \varnothing$,
- $\mathrm{pa}(2,0) := \varnothing$,
- $\forall t \in T \backslash \{0\} : \mathrm{pa}(1,t)) = \{(1,t-1),(2,t-1)\}$,
- $\forall t \in T \backslash \{0\} : \mathrm{pa}(2,t)) = \{(1,t-1),(2,t-1)\}$,

Rename:

- $\{X_{1,t}\}_{t\in T} =: \{E_t\}_{t\in T}$,
- $\{X_{2,t}\}_{t\in T} =: \{M_t\}_{t\in T}$.

The Bayesian network is called the perception-action loop, $\{E_t\}_{t\in T}$ is called the environment process and $\{M_t\}_{t\in T}$ the agent or memory process.

It is customary to explicitly represent actions and perceptions in the perception-action loop. In Definition 43 these are implicit. The interactions between agent and environment are not represented by random variables. As mentioned before in the tradition of perception action loops all interactions between agent and environment are considered to be captured by the actions and the perceptions (Beer, 1995).

We can construct sequences of random variables representing for each t the influences of $\{M_t\}_{t\in T}$ on $\{E_t\}_{t\in T}$ and vice versa such that the probability distribution p_V over all random variables remains the same. This means we introduce another Bayesian network, which we will also call the perception-action loop containing two more processes, the action process $\{A_t\}_{t\in T}$ and the sensor process $\{S_t\}_{t\in T}$. The result will be the perception-action loop considered in Bertschinger et al. (2006, 2008).

The action at a time t is a block in a partition μ_t of the state space $\mathcal{M}_t$ of the agent at t. These blocks are composed out of agent states that have the same effect on the environment's transitions from E_t to E_{t+1}. Formally, we define the partition μ_t via the equivalence relation $\equiv_{\mu_t}$ relating the elements of each block in μ_t.

Definition 44. *Given a perception-action loop $\{X_i\}_{i\in V}$. For each time $t \in T$ and $m_t^1, m_t^2 \in \mathcal{M}_t$ let*

$$m_t^1 \equiv_{\mu_t} m_t^2 \Leftrightarrow \forall e_{t+1} \in \mathcal{E}_{t+1}, e_t \in \mathcal{E}_t : p_{E_{t+1}}(e_{t+1}|m_t^1, e_t) = p_{E_{t+1}}(e_{t+1}|m_t^2, e_t). \tag{3.69}$$

Then:

(i) The action partition μ_t is then defined as the set of equivalence classes of the equivalence relation $\equiv_{\mu_t}$.

(ii) The set of actions is defined as $\mathcal{A}_t := \mu_t$ and an element $a_t \in \mathcal{A}_t$ (which is also a block in μ_t is called an action.

(iii) The action function $f_{A_t} : \mathcal{M}_t \to \mathcal{A}_t$ is defined by

$$f_{A_t}(m_t) = m_t/\mu_t, \tag{3.70}$$

where m_t/μ_t is the block in μ_t containing m_t (which is also an action).

Remark:

- The construction of the action partition[3] is not new. It is also used for example in Balduzzi (2011) to obtain coarser states (alphabet) of joint random variables. It is also similar to the construction of causal states (Shalizi, 2001). Causal states are usually a partition of pasts $x_{\leq t}$ according to equal future morphs $p(X_{t<}|x_{\leq t})$. Here we use equal transition probabilities ("transition morphs") of another process to partition the current states.

In the same way we define perceptions or sensor values via a partition ϵ_t of $\mathcal{E}_t$.

Definition 45. *Given a perception-action loop $\{X_i\}_{i\in V}$. For each time $t \in T$ and $\hat{e}_t, \bar{e}_t \in \mathcal{E}_t$ let*

$$\hat{e}_t \equiv_{\epsilon_t} \bar{e}_t \Leftrightarrow \forall m_{t+1} \in \mathcal{M}_{t+1}, m_t \in \mathcal{M}_t : p_{M_{t+1}}(m_{t+1}|m_t, \hat{e}_t) = p_{M_{t+1}}(m_{t+1}|m_t, \bar{e}_t). \tag{3.71}$$

Then:

(i) The sensor partition ϵ_t is then defined as the set of equivalence classes of the equivalence relation $\equiv_{\epsilon_t}$.

(ii) The set of sensor values is defined as $\mathcal{S}_t := \epsilon_t$ and an element $s_t \in \mathcal{S}_t$ (which is also a block in ϵ_t is called a perception of a sensor value.

(iii) The sensor function $f_{S_t} : \mathcal{E}_t \to \mathcal{S}_t$ is defined by

$$f_{S_t}(e_t) = e_t/\epsilon_t, \tag{3.72}$$

where e_t/ϵ_t is the block in ϵ_t containing e_t (which is also a sensor value).

With these definitions we can extend the Bayesian network of the perception-action loop by the action process $\{A_t\}_{t\in T}$ and the sensor process $\{S_t\}_{t\in T}$ without altering the probability distribution p_V over all random variables $\{X_i\}_{i\in V}$ in the orginal perception-action loop of Definition 43.

First, we define the extended perception-action loop. For the Bayesian network see Fig. 3.9.

Definition 46 (Extended perception-action loop)**.** *Let $\{X_i\}_{i\in V} = \{M_t, E_t\}_{t\in T}$ be a perception-action loop. Then add to $\{X_i\}_{i\in V}$ the sets of random variables $\{A_t\}_{t\in T}$ called the action process and $\{S_t\}_{t\in T}$ called the sensor process such that*

- $\forall t \in T$
- $\mathrm{pa}(M_{t+1}) := \{M_t, S_t\}$,
- $\mathrm{pa}(A_t) := M_t$,
- $\mathrm{pa}(S_t) := E_t$,
- $\mathrm{pa}(E_{t+1}) := \{E_t, A_t\}$.

[3] The author thanks Benjamin Heuer for originally pointing us to this construction.

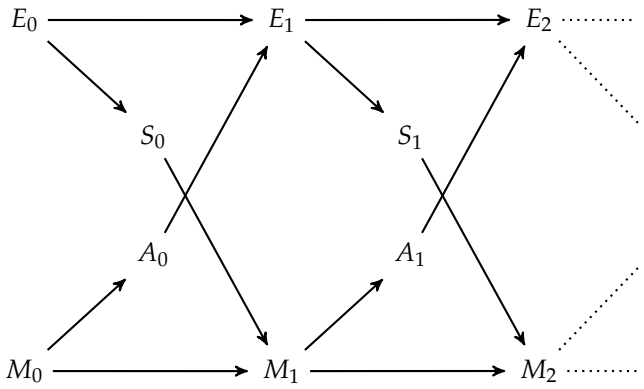


Figure 3.9: First time-steps of the Bayesian network of the extended perception-action loop. The processes $\{A_t\}_{t\in T}$ and $\{S_t\}_{t\in T}$ mediate all interactions between $\{M_t\}_{t\in T}$ and $\{E_t\}_{t\in T}$ without changing the probability distributions over the latter (see Theorem 10).

Furthermore, with f_{A_t} the action functions, f_{S_t} the sensor functions, p^w denoting probability distributions of the extended perception-action loop, and p denoting probability distributions of the original perception-action loop:

- $p^w_{M_0,E_0}(m_0, e_0) := p_{M_0,E_0}(m_0, e_0)$
- $p^w_{A_t}(a_t|m_t) := \delta_{f_{A_t}(m_t)}(a_t)$,
- $p^w_{S_t}(s_t|e_t) := \delta_{f_{S_t}(e_t)}(s_t)$,
- $p^w_{M_{t+1}}(m_{t+1}|m_t, s_t) := p_{M_{t+1}}(m_{t+1}|m_t, e_t \in f^{-1}_{S_t}(s_t))$,
- $p^w_{E_{t+1}}(e_{t+1}|a_t, e_t) := p_{E_{t+1}}(e_{t+1}|m_t \in f^{-1}_{A_t}(a_t), e_t)$.

Then the resulting Bayesian network $\{X_i\}_{i\in W} = \{M_t, A_t, S_t, E_t\}_{t\in\mathbb{N}^+}$ with probability distribution p^w_W is called the extended perception-action loop of Xv.

Remark:

- Since, by definition of f_{S_t}, for any $t \in T$ and $s_t \in \mathcal{S}_t$ all $\bar{e}_t \in f^{-1}_{S_t}(s_t)$ have the same $p_{M_{t+1}}(m_{t+1}|m_t, \bar{e}_t)$ the definition of $p^w_{M_{t+1}}(m_{t+1}|m_t, s_t)$ is unambiguous. The corresponding argument holds for f_{A_t} so that the definition of $p^w_{E_{t+1}}(e_{t+1}|a_t, e_t)$ is also unambiguous.

We then have the following theorem:

Theorem 10 (Invariant extension theorem). *Given a perception action loop $\{X_i\}_{i\in V} = \{M_t, E_t\}_{t\in T}$ and its extended perception-action loop $\{X_i\}_{i\in W} = \{M_t, A_t, S_t, E_t\}_{t\in\mathbb{N}^+}$. Let $p_V = p_{M_T,E_T}$ be the probability distribution over the entire perception action loop $\{X_i\}_{i\in V}$ and let $p^w_{M_T,E_T}$ be the marginal probability distribution over the memory and environment process obtained from the probability distribution p^w_W over the entire extended perception-action loop. Then*

$$p_{M_T,E_T} = p^w_{M_T,E_T}. \tag{3.73}$$

Proof.

$$p^w_{M_T,E_T}(m_T,e_T) = \sum_{a_T}\sum_{s_T}\prod_{t=1}^{n-1} p^w_{M_t}(m_t|m_{t-1},s_{t-1})p^w_{S_{t-1}}(s_{t-1}|e_{t-1})$$
$$p^w_{E_t}(e_t|a_{t-1},e_{t-1})p^w_{A_{t-1}}(a_{t-1}|m_{t-1})p^w_{M_0,E_0}(m_0,e_0) \tag{3.74}$$

$$= \sum_{a_T}\sum_{s_T}\prod_{t=1}^{n-1} p^w_{M_t}(m_t|m_{t-1},s_{t-1})\delta_{f_{S_{t-1}}(e_{t-1})}(s_{t-1})$$
$$p^w_{E_t}(e_t|a_{t-1},e_{t-1})\delta_{f_{A_{t-1}}(m_{t-1})}(a_{t-1})p^w_{M_0,E_0}(m_0,e_0) \tag{3.75}$$

$$= \prod_{t=1}^{n-1} p^w_{M_t}(m_t|m_{t-1},f_{S_{t-1}}(e_{t-1}))$$
$$p^w_{E_t}(e_t|f_{A_{t-1}}(m_{t-1}),e_{t-1})p^w_{M_0,E_0}(m_0,e_0) \tag{3.76}$$

$$= \prod_{t=1}^{n-1} p_{M_t}(m_t|m_{t-1},\bar{e}_{t-1}\in f^{-1}_{S_{t-1}}\circ f_{S_{t-1}}(e_{t-1}))$$
$$p_{E_t}(e_t|\bar{m}_{t-1}\in f^{-1}_{A_{t-1}}\circ f_{A_{t-1}}(m_{t-1}),e_{t-1})p_{M_0,E_0}(m_0,e_0) \tag{3.77}$$

$$= p_{M_T,E_T}(m_T,e_T). \tag{3.78}$$

□

Remarks:

- This proof shows that the introduction of action and sensor process in the way described by Definition 46 only makes the interactions between agent and environment processes explicit. Action and sensor processes are not essential to the perception-action loop and do not introduce any additional dynamics. They only represent what the environment "sees" of the agent and vice versa. In other words the dynamics of agent and environment do not *require* that the states space $\mathcal{M}_t \times \mathcal{E}_t$ of a time-slice is extended to $\mathcal{M}_t \times \mathcal{A}_t \times \mathcal{S}_t \times E_t$. This insight is also in line with Bertschinger et al. (2006) which uses the extended perception-action loop and refers to the action and sensor processes as "channels".

- This proof also shows that the sensor process (and conversely the actions) captures all influences from the environment on the agent. Else the dynamics of the extended perception-action loop could remain identical. We use this fact as a starting point for our conception of entity

perception in Section 5.5. There we want to capture all influences of the environment on a set of "spatiotemporal patterns" or "entities" instead of on a stochastic process like $\{M_t\}_{t \in T}$. This will require a generalization of the perception extraction procedure in Definition 45.

Chapter 4

Spatiotemporal patterns

This chapter constitutes the formal part of this thesis. We present here a first investigation of the properties and structure of spatiotemporal patterns[1] (STP) in Bayesian networks in relation to the novel measure of *specific local integration* (SLI). Apart from SLI we also present the derived measure of complete local integration (CLI). The main result is the *disintegration theorem* which relates the SLI of whole trajectories of Bayesian networks to the CLI of parts of these trajectories and vice versa. The connection between the two is revealed by the *disintegration hierarchy* (Definition 54) and its *refinement-free* version (Definition 56). These are our own constructions. In Section 4.5 we also formally define the effect of symmetry transformations on STPs in Bayesian networks and derive the behaviour of SLI under such transformations. The main result in this respect are the SLI symmetry theorems (Theorems 30 and 31). Finally, we establish under what circumstances spatial symmetries spread throughout the entire Bayesian network if the Bayesian networks are multivariate Markov chains or driven multivariate Markov chains. The according theorems are not new but provide the connection to more practical scenarios of cellular automata and reaction diffusion systems. Later, in Chapter 6 we visualise the disintegration hierarchies, completely locally integrated STPs, and use the SLI symmetry theorems to explain its structure.

In more detail the chapter contains the following:

- In Section 4.1 we give the definition of the partition lattice of Bayesian networks. This lattice is an underlying structure throughout this thesis.
- In Section 4.2 we define SLI. We constructively prove its upper bounds and construct an example of a STP with a particularly low (and negative) SLI. These constructions are mainly of technical interest and not employed conceptually. We also propose a normalised version of SLI which will not be further used but may be of interest for future research.

[1] In this section we will always speak of spatiotemporal patterns in Bayesian networks instead of patterns in sets of random variables. There is no formal difference however. We only want to emphasise that we will consider spatiotemporally extended patterns whenever there are well defined notions of space and time in this thesis. Also note that every set of random variables can just be seen as a Bayesian network (possibly without any edges).

Finally we derive some algebraic properties of differences between specific local integrations. These are given for technical reference.

- In Section 4.3 we state the definition of CLI. This is an important notion throughout this thesis and the basis for the definition of entities in multivariate Markov chains in Section 5.3.
- In Section 4.4 we define the disintegration hierarchies and prove the disintegration theorem.
- In Section 4.5 we first introduce notation and terminology to express the effect of permutations of nodes within a Bayesian network on STPs, partitions, and probabilities of spatiotemporal patterns. Then we prove the SLI symmetry theorems which specify the behaviour of SLI under permutation symmetries of the Bayesian network. These can provide insights into the structure of the disintegration hierarchies for systems with high degrees of symmetries like cellular automata. We will see how this can be done in Chapter 6. Furthermore, symmetry properties are of general interest for future theoretical developments.
- In Section 4.6 we look at the Bayesian networks that are multivariate Markov chains and derive conditions under which spatial symmetries of initial distribution and Markov matrix spread the spatial symmetry over the entire Bayesian network. We also do this for driven multivariate Markov chains. The SLI symmetry theorems concern symmetries of STPs i.e. they depend on symmetries that are not purely spatial. The theorems in this section provides a way to obtain such extended symmetries from more simple and often well known ones. The inclusion of driven multivariate Markov chains also extends the applicability of our formal results beyond cellular automata to driven systems.

In summary the original contributions of this chapter are:

- Definition of SLI.
- Constructive proof of upper bound of SLI.
- Construction of negative SLI example.
- Definition of CLI.
- Definition of disintegration hierarchy and refinement-free disintegration hierarchy.
- Proof of the disintegration theorem.
- Proof of the SLI symmetry theorems.

4.1 Partition lattice of Bayesian networks

In this section we introduce the partition lattice of a Bayesian network which is the underlying structure for most of the developments in this chapter. All STPs occupy blocks of partitions in this lattice and can be further partitioned into blocks that are also blocks in the partition lattice of the Bayesian network. We also introduce the set of anti-STPs with respect to a partition which is a generalisation of the set of anti-patterns $\neg(x_A)$ (Definition 20) of a STP x_A. This will be mainly used in proofs and is not important for the rest of this thesis.

Definition 47 (Partition lattice of a Bayesian network)**.** *Given a Bayesian network $\{X_i\}_{i\in V}$ and a subset $A \subseteq V$ we denote the partition lattice of A by $\mathfrak{L}(A)$. Every partition $\pi \in \mathfrak{L}(A)$ also naturally induces the following partitions:*

(i) partition $\pi(X_A)$ of the joint random variable X_A by defining for every block $b \in \pi$ the corresponding block $X_b \in \pi(X_A)$
(ii) partition $\pi(x_A)$ of every STP $x_A \in X_A$ by defining for each block $b \in \pi$ the corresponding block $x_b \in \pi(x_A)$.

When it will be clear from context which instance of π we are referring to we will denote these two partitions in the following also just by π.

Remark:

- Partition lattices were defined in Definition 9
- $\mathfrak{L}(V)$ is the partition lattice of all nodes in the Bayesian network.
- The partition lattice $\mathfrak{L}(A)$ of the index set A or equivalently the nodes in the Bayesian network indexed by A must not be confused with the partition lattice $\mathfrak{L}(\mathcal{X}_A)$ of the state space $\mathcal{X}_A$ of the joint random variable of those nodes.

Definition 48. *Given a Bayesian network $\{X_i\}_{i\in V}$, an STP x_A and a partition $\pi \in \mathfrak{L}(A)$ of x_A the set of anti-STPs of x_A with respect to π denoted by $\neg_\pi(x_A)$ is defined via*

$$\neg_\pi(x_A) := \{\bar{x}_A \in \mathcal{X}_A : \forall b \in \pi, \bar{x}_b \neq x_b\}. \tag{4.1}$$

Remark:

- Note that $\neg(x_A) = \neg_{\mathbf{0}}(x_A)$. Recall that $\mathbf{0}$ is the finest partition in a partition lattice and contains only blocks that are singletons (see Definition 4).

Theorem 11. *Given a Bayesian network $\{X_i\}_{i\in V}$, an STP x_A and two partitions ξ, π of x_A with $\xi \lhd \pi$ we have:*

$$\neg_\xi(x_A) \subseteq \neg_\pi(x_A). \tag{4.2}$$

Proof.

$$\begin{aligned}
\neg_\xi(x_A) :&= \{\bar{x}_A \in \mathcal{X}_A : \forall b \in \xi, \bar{x}_b \neq x_b\} && (4.3)\\
&= \{\bar{x}_A \in \mathcal{X}_A : \forall b \in \xi, \exists i \in b, \bar{x}_i \neq x_i\} && (4.4)\\
&= \{\bar{x}_A \in \mathcal{X}_A : \forall c \in \pi, \forall b \in \xi \text{ with } b \subseteq c, \exists i \in b, \bar{x}_i \neq x_i\} && (4.5)\\
&\subseteq \{\bar{x}_A \in \mathcal{X}_A : \forall c \in \pi, \exists i \in c, \bar{x}_i \neq x_i\} && (4.6)\\
&= \{\bar{x}_A \in \mathcal{X}_A : \forall c \in \pi, \bar{x}_c \neq x_c\} && (4.7)\\
&= \neg_\pi(x_A). && (4.8)
\end{aligned}$$

□

Remark:

- This theorem shows that the anti-STP $\neg_\xi(x_A)$ of a partition ξ that refines a partition π is the more restrictive anti-STP. It requires for at least one variable $\bar{x}_i$ of each block b that it differs from x_i whereas the anti-STPs with respect to the coarser partition $\pi \rhd \xi$ only requires one such differing value across all the blocks of ξ that make up each block of π.

4.2 Specific local integration

This section introduces the specific local integration (SLI). It also proves its upper bounds constructively and constructs an example of negative SLI. We state a definition of normalised SLI and algebraic properties of differences between specific local integrations.

4.2.1 General and deterministic case

Definition 49 (Specific local integration (SLI))**.** *Given a Bayesian network $\{X\}_{i\in V}$ and a STP x_O the specific local integration $\mathrm{mi}_\pi(x_O)$ of x_O with respect to a partition π of $O \subseteq V$ is defined as*

$$\mathrm{mi}_\pi(x_O) := \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)}. \tag{4.9}$$

In this thesis we use the convention that $\log \frac{0}{0} := 0$.

Theorem 12 (Deterministic specific local integration)**.** *Given a deterministic Bayesian network (Definition 32) and uniform initial distribution the SLI of x_O with respect to partition π can be expressed in another way: Let $N(x_O)$ refer to the number of trajectories in which x_O occurs. Then*

$$\mathrm{mi}_\pi(x_O) = (|\pi| - 1)\log|\mathcal{X}_{V_0}| - \log\frac{\prod_{b\in\pi} N(x_b)}{N(x_O)}. \tag{4.10}$$

Proof. Follows by replacing the probabilities $p_O(x_O)$ and $p_b(x_b)$ in Eq. (4.9) with their deterministic expressions from Theorem 8, i.e. $p_A(x_A) = N(X_A)/|\mathcal{X}_{V_0}|$.

Then:

$$\mathrm{mi}_\pi(x_O) := \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)} \tag{4.11}$$

$$= \log \frac{\frac{N(x_O)}{|\mathcal{X}_{V_0}|}}{\prod_{b\in\pi} \frac{N(x_b)}{|\mathcal{X}_{V_0}|}} \tag{4.12}$$

$$= \log \frac{\frac{N(x_O)}{|\mathcal{X}_{V_0}|}}{|\mathcal{X}_{V_0}|^{-|\pi|} \prod_{b\in\pi} N(x_b)} \tag{4.13}$$

$$= \log \frac{|\mathcal{X}_{V_0}|^{|\pi|-1} N(x_O)}{\prod_{b\in\pi} N(x_b)} \tag{4.14}$$

$$= (|\pi|-1)\log|\mathcal{X}_{V_0}| - \log \frac{\prod_{b\in\pi} N(x_b)}{N(x_O)}. \tag{4.15}$$

□

4.2.2 Upper bounds

In this section we prove upper bounds of SLI. It is not essential for the rest of the thesis and is presented mainly for technical reference since bounds are important aspects of measures. However it is also useful to familiarise the reader with the measure of SLI since we prove the bounds constructively. We first show constructively that if we can choose the Bayesian network and the STP then SLI can be arbitrary large. This construction sets the probabilities of all blocks equal to the probability of the STP. In the subsequent theorem we show that this property in general gives the upper bound of SLI if the cardinality of the partition is fixed. This leads directly tho the upper bound if the cardinality of the partition is not fixed in the next theorem. Finally we give the expressions of the bounds in the deterministic case for convenient reference.

Theorem 13 (Construction of a STP with maximum SLI)**.** *Given a probability $q \in (0,1)$ and a positive natural number n we can construct a Bayesian network $\{X_i\}_{i\in V}$ and an STP x_O such that*

$$\mathrm{mi}_\pi(x_O) = -(n-1)\log q. \tag{4.16}$$

Proof. We construct a Bayesian network which realises two conditions on the probability p_O. From these two conditions (which can also be realised by other Bayesian networks) we can then derive the theorem.

Choose a Bayesian network $\{X_i\}_{i\in V}$ with binary random variables $\mathcal{X}_i = \{0,1\}$ for all $i \in V$. Choose all nodes in O dependent only on node $j \in O$, the dependence of the nodes in $V\backslash O$ is arbitrary:

- for all $i \in O \subset V$ let $\mathrm{pa}(i) \cap (V\backslash O) = \varnothing$, i.e. nodes in O have no parents in the complement of O,
- for a specific $j \in O$ and all other $i \in O\backslash\{j\}$ let $\mathrm{pa}(i) = \{j\}$, i.e. all nodes in O apart from j have $j \in O$ as a parent,

- for all $i \in O \backslash \{j\}$ let $p_i(\bar{x}_i | b\bar{x}_j) = \delta_{\bar{x}_j}(\bar{x}_i)$, i.e. the state of all nodes in O is always the same as the state of node j,
- also choose $p_j(x_j) = q$ and $\sum_{\bar{x}_j \neq x_j} p_j(x_j) = 1 - q$.

Then it is straightforward to see that:

1. $p_O(x_O) = q$,
2. $\sum_{\bar{x}_O \in \neg(x_O)} p_O(\bar{x}_O) = 1 - q$.

Note that there are many Bayesian networks that realise the latter two conditions for some x_O. These latter two conditions are the only requirements for the following calculation.

Next note that the two conditions imply that $p_O(\bar{x}_O) = 0$ if neither $\bar{x}_O = x_O$ nor $\bar{x}_O \in \neg(x_O)$. Then for every partition π of O with $|\pi| = n$ and $n > 1$ we have

$$\begin{aligned}
\mathrm{mi}_\pi(x_O) &= \log \frac{p_O(x_O)}{\prod_{b \in \pi} p_b(x_b)} && (4.17)\\
&= \log \frac{p_O(x_O)}{\prod_{b \in \pi} \sum_{\bar{x}_{O \backslash b}} p_O(x_b, \bar{x}_{O \backslash b})} && (4.18)\\
&= \log \frac{p_O(x_O)}{\prod_{b \in \pi} \left(p_O(x_O) + \sum_{\bar{x}_{O \backslash b} \neq x_{O \backslash b}} p_O(x_b, \bar{x}_{O \backslash b}) \right)} && (4.19)\\
&= \log \frac{p_O(x_O)}{\prod_{b \in \pi} p_O(x_O)} && (4.20)\\
&= \log \frac{p_O(x_O)}{p_O(x_O)^n} && (4.21)\\
&= -(n-1) \log q. && (4.22)
\end{aligned}$$

□

Remark:

- We will use this construction to reveal the general tight upper bound of $\mathrm{mi}_\pi(x_O)$.
- The construction used here ensures that the probability $p_b(x_b)$ of each block $b \in \pi$ is equal to the probability of the STP $p_O(x_O) = q$. In other words, the parts of x_O that are indicated by π all occur if and only if the whole STP x_O occurs. Note that in general x_b always occurs if x_O occurs but not vice versa.

Theorem 14 (Upper bound of SLI)**.** *For any Bayesian network $\{X\}_{i \in V}$ and STP x_O*

(i) The tight upper bound of the SLI with respect to partition π is

$$\mathrm{mi}_\pi(x_O) \leqslant -(|\pi| - 1) \log p_O(x_O). \tag{4.23}$$

(ii) The upper bound is achieved if and only if for all $b \in \pi$ we have

$$p_b(x_b) = p_O(x_O). \tag{4.24}$$

(iii) The upper bound is achieved if and only if for all $b \in \pi$ we have that x_O occurs if and only if x_b occurs.

Proof. **ad (i)** By Definition 49 we have

$$\mathrm{mi}_\pi(x_O) = \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)}. \tag{4.25}$$

Now note that for any x_O and $b \subseteq O$

$$\begin{aligned} p_b(x_b) &= \sum_{\bar{x}_{O\setminus b}} p_O(x_b, \bar{x}_{O\setminus b}) && (4.26)\\ &= p_O(x_O) + \sum_{\bar{x}_{O\setminus b} \neq x_{O\setminus b}} p_O(x_b, \bar{x}_{O\setminus b}) && (4.27)\\ &\geqslant p_O(x_O). && (4.28) \end{aligned}$$

Plugging this into Eq. (4.25) for every $p_b(x_b)$ we get

$$\begin{aligned} \mathrm{mi}_\pi(x_O) &= \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)} && (4.29)\\ &\leqslant \log \frac{p_O(x_O)}{p_O(x_O)^{|\pi|}} && (4.30)\\ &= -(|\pi| - 1) \log p_O(x_O). && (4.31) \end{aligned}$$

This shows that $-(|\pi| - 1) \log p_O(x_O)$ is indeed an upper bound. To show that it is tight we have to show that for a given $p_O(x_O)$ and $|\pi|$ there are Bayesian networks with STPs x_O such that this upper bound is achieved. The construction of such a Bayesian network and an STP x_O was presented in Theorem 13.

ad (ii)) If for all $b \in \pi$ we have $p_b(x_b) = p_O(x_O)$ then clearly $\mathrm{mi}_\pi(x_O) = -(|\pi| - 1) \log p_O(x_O)$ and the least upper bound is achieved. If on the other hand $\mathrm{mi}_\pi(x_O) = -(|\pi| - 1) \log p_O(x_O)$ then

$$\begin{aligned} & \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)} = -(|\pi| - 1) \log p_O(x_O) && (4.32)\\ \Leftrightarrow \quad & \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)} = \log \frac{p_O(x_O)}{p_O(x_O)^{|\pi|}} && (4.33)\\ \Leftrightarrow \quad & \prod_{b\in\pi} p_b(x_b) = p_O(x_O)^{|\pi|}, && (4.34) \end{aligned}$$

and because $p_b(x_b) \geqslant p_O(x_O)$ (Eq. (4.28)) any deviation of any of the $p_b(x_b)$ from $p_O(x_O)$ leads to $\prod_{b\in\pi} p_b(x_b) > p_O(x_O)^{|\pi|}$ such that for all $b \in \pi$ we must have $p_b(x_b) = p_O(x_O)$.

ad (iii) By definition for any $b \in \pi$ we have $b \subseteq O$ such that x_b always occurs if x_O occurs. Now assume x_b occurs and x_O does not occur. In that case there is a positive probability for an STP $(x_b, \bar{x}_{O\setminus b})$ with $\bar{x}_{O\setminus b} \neq x_{O\setminus b}$ i.e. $p_O(x_b, \bar{x}_{O\setminus b}) > 0$. Recalling Eq. (4.27) we then see that

$$p_b(x_b) = p_O(x_O) + \sum_{\bar{x}_{O\setminus b} \neq x_{O\setminus b}} p_O(x_b, \bar{x}_{O\setminus b}) \tag{4.35}$$

$$> p_O(x_O). \tag{4.36}$$

which contradicts the fact that $p_b(x_b) = p_O(x_O)$ so x_b cannot occur without x_O occurring as well.

□

Remarks:

- Note that this is the least upper bound for Bayesian networks in general. For a specific Bayesian network there might be no STP that achieves this bound.
- So the least upper bound of SLI $\mathrm{mi}_\pi(x_O)$ is the self-information $-\log p(x_O)$ of the STP x_O multiplied by one less than the cardinality $|\pi|$ of the partition.
- In other words, the maximally possible SLI increases with the improbability of the STP and the number of parts that it is split into.
- For an STP x_O that achieves the least upper bound of SLI, the occurrence of any part x_b indicated by the partition π of O implies the occurrence of the entire STP.
- Using this least upper bound it is easy to derive a least upper bound for the SLI of an STP x_O across all partitions $|\pi|$. We just have to note that $|\pi| \leqslant |O|$. This leads directly to the next statement.
- Further down we will use the least upper bound for specific partitions in order to normalise the SLI.

Theorem 15. *For any Bayesian network $\{X_i\}_{i\in V}$ and STP x_O the least upper bound of the SLI with respect to arbitrary partitions is*

$$\max_\pi \mathrm{mi}_\pi(x_0) \leqslant -(|O| - 1) \log p_O(x_O). \tag{4.37}$$

Proof. Follows from Theorem 14 and the fact that for an arbitrary partition π of O we have $|\pi| \leqslant |O|$. □

Theorem 16 (Deterministic least upper bounds)**.** *For any deterministic Bayesian network (Definition 32) $\{X_i\}_{i\in V}$ with uniform initial distribution p_{V_0} we find the following bounds.*

(i) The least upper bound of the SLI with respect to partition π for a given STP x_O is

$$\mathrm{mi}_\pi(x_O) \leqslant -(|\pi| - 1) \log \frac{N(x_O)}{|\mathcal{X}_{V_0}|} \tag{4.38}$$

(ii) *The least upper bound of the SLI for an STP x_O across all partitions is*

$$\max_{\pi} \mathrm{mi}_\pi(x_0) \leqslant -(|O|-1)\log\frac{N(x_O)}{|\mathcal{X}_{V_0}|} \tag{4.39}$$

(iii) *The least upper bound of the SLI with respect to partitions π of cardinality $|\pi|$ across all STP x_O is*

$$\max_{x_O} \mathrm{mi}_\pi(x_O) \leqslant (|\pi|-1)\log|\mathcal{X}_{V_0}| \tag{4.40}$$

(iv) *The least upper bound of the SLI across all partitions and all STP is*

$$\max_{\pi}\max_{x_O} \mathrm{mi}_\pi(x_O) \leqslant (|V|-1)\log|\mathcal{X}_{V_0}| \tag{4.41}$$

Proof. **ad (i)** Follows directly from Theorem 14 by replacing $p_O(x_O)$ with $\frac{N(x_O)}{|\mathcal{X}_{V_O}|}$ as shown in Eq. (3.41).

ad (ii) Follows from Theorem 16.(i) and $|\pi| \leqslant |O|$ for all partitions π of $|O|$.

ad (iii) Follows from Theorem 16.(i) and $N(x_O) \geqslant 1$ for all x_O.

ad (iv) Follows from Theorem 16.(iii) and $|\pi| \leqslant |V|$ for all partitions π of all $O \subseteq V$.

□

Remarks:

- Again these are tight upper bounds among all deterministic Bayesian networks with uniform initial distribution. Particular instances of such Bayesian networks may not contain any STP which achieve these bounds.
- Note that the first term in Eq. (4.10) corresponds to the tight upper bound in Theorem 16.(iii). So $\mathrm{mi}_\pi(x_O)$ is just this upper bound minus the second term which is always positive as $\prod_b N(x_b) \geqslant N(x_O)$.

4.2.3 Negative SLI

This section shows that SLI of an STP x_O with respect to partition π can be negative *independent* of the the probability of x_O (as long as it is not 1) and the cardinality of the partition (as long as that is not 1). This is not important for the rest of the thesis but is of technical interest in its own right. It also shows how to get negative SLI at all which may not be obvious.

Theorem 17. *For any given probability $q < 1$ and cardinality $|\pi| > 1$ of a partition π there exists an STP x_O in a Bayesian network $\{X_i\}_{i\in V}$ such that $q = p_O(x_O)$ and*

$$\mathrm{mi}_\pi(x_O) < 0. \tag{4.42}$$

Proof. We construct the probability distribution $p_O : \mathcal{X}_O \to [0,1]$ and ignore the behaviour of the Bayesian network $\{X_i\}_{i\in V}$ outside of $O \subseteq V$. In any case $\{X_i\}_{i\in O}$ is also by itself a Bayesian network. We define (see remarks below for some intuitions behind these definitions):

(i) for all $i \in O$ let $|\mathcal{X}_i| = n$

(ii) for every block $b \in \pi$ let $|b| = \frac{|O|}{|\pi|}$,
(iii) for $\bar{x}_O \in \mathcal{X}_O$ let:

$$p_O(\bar{x}_O) := \begin{cases} q & \text{if } \bar{x}_O = x_O, \\ \frac{1-q-d}{\sum_{b\in\pi} |\neg(x_b)|} & \text{if } \exists c \in \pi \text{ s.t. } \bar{x}_{O\setminus c} = x_{O\setminus c} \wedge \bar{x}_c \neq x_c, \\ \frac{d}{|\neg(x_O)|} & \text{if } \bar{x}_O \in \neg(x_O), \\ 0 & \text{else.} \end{cases} \tag{4.43}$$

Then we can calculate the SLI. First note that according to (i) and (ii) we have $|\mathcal{X}_b| = |\mathcal{X}_c|$ for all $b, c \in \pi$ and therefore also $|\neg(x_b)| = |\neg(x_c)|$ for all $b, c \in \pi$. So let $m := |\neg(x_b)|$. Then note that according to (iii) for all $b \in \pi$

$$\sum_{\bar{x}_{O\setminus b} \neq x_{O\setminus b}} p_O(x_b, \bar{x}_{O\setminus b}) = \sum_{c\in\pi\setminus b} \sum_{\bar{x}_c \neq x_c} p_O(x_b, x_{O\setminus(b\cup c)}, \bar{x}_c) \tag{4.44}$$

$$= \sum_{c\in\pi\setminus b} \sum_{\bar{x}_c \neq x_c} \frac{1-q-d}{\sum_{b\in\pi} |\neg(x_b)|} \tag{4.45}$$

$$= \sum_{c\in\pi\setminus b} \sum_{\bar{x}_c \neq x_c} \frac{1-q-d}{m|\pi|} \tag{4.46}$$

$$= \sum_{c\in\pi\setminus b} \frac{1-q-d}{m|\pi|} |\neg(x_c)| \tag{4.47}$$

$$= \frac{|\pi|-1}{|\pi|}(1-q-d) \tag{4.48}$$

Plug this into the SLI definition:

$$\mathrm{mi}_\pi(x_O) = \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)} \tag{4.49}$$

$$= \log \frac{q}{\prod_{b\in\pi} q + \sum_{\bar{x}_{O\setminus b} \neq x_{O\setminus b}} p_O(x_b, \bar{x}_{O\setminus b})} \tag{4.50}$$

$$= \log \frac{q}{\prod_{b\in\pi} q + \frac{|\pi|-1}{|\pi|}(1-q-d)} \tag{4.51}$$

$$= \log \frac{q}{\left(1 - \frac{1-q}{|\pi|}\right)^{|\pi|}}. \tag{4.52}$$

If we now set $d = 0$ then we can use Bernoulli's inequality[2] to prove that this is negative for $0 < q < 1$ and $|\pi| \geqslant 2$. Bernoulli's inequality is

$$(1+x)^n \geqslant 1 + nx \tag{4.53}$$

for $x \geqslant -1$ and n a natural number. Replacing x by $-(1-q)/|\pi|$ we see that

$$\left(1 - \frac{1-q}{|\pi|}\right)^{|\pi|} > q \tag{4.54}$$

[2]We thank von Eitzen (2016) for pointing this out. An example reference for Bernoulli's inequality is Bullen (2003).

such that the argument of the logarithm is smaller than one which gives us negative SLI.

□

Remarks:

- The construction used to proof this theorem with the achieved value in Eq. (4.52) is also our best candidate for a tight lower bound of SLI for a given $p_O(x_O)$ and $|\pi|$. However, we have not been able to prove this yet.
- The construction equidistributes the probability $1-q$ (left to be distributed after q is chosen) to the STPs $\bar{x}_O$ that are *almost* the same as the STP x_O. They are almost the same in a precise sense because they only differ in only one of the blocks of π they differ by as little as can possibly be resolved/revealed by the partition π.
- In order to achieve the negative SLI of Eq. (4.52) the requirement is only that Eq. (4.48) is satisfied. Our construction shows one way how this can be achieved.
- For a pattern and partition such that $|O|/|\pi|$ is not a natural number, the same bound might still be achieved however a little extra effort has to go into the construction (iii) such that Eq. (4.48) still holds. This is not necessary for our purpose here as we only want to show the existence of patterns that with negative SLI.
- An interpretation of the construction is that STPs which either occur as a whole or (with uniform probability) missing exactly one part always have negative SLI.

4.2.4 Normalised specific local integration

Here we present a way to employ the least upper bound to define a normalised version of SLI. This notion is not important for the rest of this thesis.

Definition 50 (Normalised specific local integration)**.** *The normalised specific local integration is just the specific normalised integration* $\mathrm{mi}_\pi(x_O)$ *divided by the least upper bound for the STP* x_O *and the partition* π*:*

$$\mathrm{nmi}_\pi(x_O) := \frac{\mathrm{mi}_\pi(x_O)}{-(|\pi|-1)\log p_O(x_O)}. \tag{4.55}$$

The value of $\mathrm{nmi}_\pi(x_O)$ *is in the the interval* $(-\infty, 1]$.

Theorem 18. *For any Bayesian network, any STP* x_O *and any partition* π

$$\mathrm{nmi}_\pi(x_O) \leqslant 1. \tag{4.56}$$

Proof. Follows from Theorem 14 and the definition of $\mathrm{nmi}_\pi(x_O)$. □

Remarks:

- The normalised SLI gives us a measure of integration which is independent of the cardinality of the partition. This means we can compare the specific local integrations of a STP across partitions π of different cardinalities.
- At the same time the normalised SLI also compensates for differences in the self-information $-\log p_O(x_O)$ of the STP x_O. This self-information can be interpreted as a measure of the "size" of the STP. For a given STP x_O this size of course does not change across the partitions. However, dividing by it allows to also compare the degree of SLI across STPs of differing sizes.

4.2.5 Difference of SLI with respect to different partitions

This section present some algebraic properties of SLI. An alternative expression for SLI, the difference between the SLI with respect to different partitions of the same STP, and between the SLI with respect to different partitions of different STP. It is presented here for reference and used only in some of the proofs that follow.

Definition 51. *Given a partition π of a set V and a subset $A \subseteq V$ we define the restricted partition $\pi|_A$ of π to A via:*

$$\pi|_A := \{b \cap A : b \in \pi\}. \tag{4.57}$$

Conversely, if ξ is a partition of $A \subseteq V$ and π a partition of V and $\pi|_A = \xi$ we call π an extension of ξ to V.

Theorem 19. *Given any particular total order $(b_1, ..., b_n)$ of the blocks of a partition π of $O \subseteq V$ with $|\pi| = n$ we can always write the SLI as a sum over the blocks in the partition:*

$$\mathrm{mi}_\pi(x_O) = \sum_{i=1}^{n-1} \log \frac{p(x_{b_{i+1}}|x_{\bigcup_{j=1}^{i} b_j})}{p_{b_{i+1}}(x_{b_{i+1}})} \tag{4.58}$$

Proof. Follows directly from the chain rule of probability and the properties of the logarithm. □

Definition 52. *Given two partitions π, ξ of $O \subseteq V$ and an STP x_O then we define the difference $\Delta\mathrm{mi}^\pi_\xi(x_O)$ of the respective specific local integrations via:*

$$\Delta\mathrm{mi}^\pi_\xi(x_O) := \mathrm{mi}_\pi(x_O) - \mathrm{mi}_\xi(x_O). \tag{4.59}$$

Theorem 20. *Given three partitions π, ξ, ρ of $O \subseteq V$ and an STP x_O. Then:*

(i)

$$\Delta\mathrm{mi}^\pi_\xi(x_O) = \log \frac{\prod_{a\in\xi} p_a(x_a)}{\prod_{b\in\pi} p_b(x_b)}. \tag{4.60}$$

(ii)

$$\Delta \mathrm{mi}_{\xi}^{\pi}(x_O) = \sum_{b \in \pi} \log \frac{1}{p_b(x_b)} - \sum_{a \in \xi} \log \frac{1}{p_a(x_a)}. \tag{4.61}$$

(iii) If $\pi \preccurlyeq \xi$ then:

$$\Delta \mathrm{mi}_{\xi}^{\pi}(x_O) = \sum_{a \in \xi} \mathrm{mi}_{\pi|_a}(x_a). \tag{4.62}$$

(iv) If $\rho \preccurlyeq \pi$ and $\rho \preccurlyeq \xi$ i.e. ρ is a lower bound of π and ξ then

$$\Delta \mathrm{mi}_{\xi}^{\pi}(x_O) = \sum_{a \in \xi} \mathrm{mi}_{\rho|_a}(x_a) - \sum_{b \in \pi} \mathrm{mi}_{\rho|_b}(x_b) \tag{4.63}$$

$$= \Delta \mathrm{mi}_{\xi}^{\rho}(x_O) - \Delta \mathrm{mi}_{\pi}^{\rho}(x_O). \tag{4.64}$$

Proof. Follows straightforwardly from the definitions and properties of the logarithm. □

Remarks:

- Theorem 20.(ii) says that the difference between the specific local integrations $\Delta \mathrm{mi}_{\xi}^{\pi}(x_O)$ of two arbitrary partitions π, ξ is equal to the difference of the according sums over the self-informations of the blocks in each partition.
- Theorem 20.(iii) says that the difference between the specific local integrations $\Delta \mathrm{mi}_{\xi}^{\pi}(x_O)$ of a refinement π of a partition ξ and the partition ξ itself is the sum over the specific local integrations $\mathrm{mi}_{\pi|_a}(x_a)$ of each of the blocks $a \in \xi$ of the original partition ξ with respect to their refinement $\pi|_a$ due to π.
- Theorem 20.(iv) says that the difference between the specific local integrations $\Delta \mathrm{mi}_{\xi}^{\pi}(x_O)$ of two arbitrary partitions π, ξ is also equal to the negative difference between the sums over the SLI of each of their blocks with respect to a partition ρ that refines both π and ξ.

Theorem 21. *Given $S \subseteq O \subseteq V$ as well as a partition π of O and a partition ξ of S we have:*

(i)

$$\mathrm{mi}_{\pi}(x_O) = \log \frac{p(x_{O \setminus S}|x_S)}{\prod_{b \in \pi} p(x_{b \setminus S}|x_{b \cap S})} + \mathrm{mi}_{\xi}(x_S) + \Delta \mathrm{mi}_{\xi}^{\pi|_S}(x_S). \tag{4.65}$$

(ii) If we set $\xi = \pi|_S$ then:

$$\mathrm{mi}_{\pi}(x_O) = \log \frac{p(x_{O \setminus S}|x_S)}{\prod_{b \in \pi} p(x_{b \setminus S}|x_{b \cap S})} + \mathrm{mi}_{\pi|_S}(x_S). \tag{4.66}$$

Proof. We will use here that $p(x_{\varnothing}) = p(x_{\varnothing}|x_O) = p(\varnothing) = 1$ and $p(x_O|x_{\varnothing}) = p(x_O)$ for any set $O \in V$ and STP x_O. This is in accordance with probability

theory. Then for (i):

$$\begin{aligned}
\mathrm{mi}_\pi(x_O) &= \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)} && (4.67)\\
&= \log \frac{p(x_{O\setminus S}|x_S)p_S(x_S)}{\prod_{b\in\pi} p(x_{b\setminus S}|x_{b\cap S})p_{b\cap S}(x_{b\cap S})} && (4.68)\\
&= \log \frac{p(x_{O\setminus S}|x_S)}{\prod_{b\in\pi} p(x_{b\setminus S}|x_{b\cap S})} + \log \frac{p_S(x_S)}{\prod_{b\in\pi} p_{b\cap S}(x_{b\cap S})} && (4.69)\\
&= \log \frac{p(x_{O\setminus S}|x_S)}{\prod_{b\in\pi} p(x_{b\setminus S}|x_{b\cap S})} + \log \frac{p_S(x_S)}{\prod_{c\in\pi|_S} p_c(x_c)} && (4.70)\\
&= \log \frac{p(x_{O\setminus S}|x_S)}{\prod_{b\in\pi} p(x_{b\setminus S}|x_{b\cap S})} + \mathrm{mi}_\xi(x_S) + \Delta\mathrm{mi}_\xi^{\pi|_S}(x_S). && (4.71)
\end{aligned}$$

Then (ii) follows by setting $\xi = \pi|_S$. □

Remarks:

- Theorem 20 follows from Theorem 21 by setting $S = O$.
- Theorem 21.(ii) says that an STP x_S with vanishing or negative $\mathrm{mi}_\xi(x_S)$ can be part of an STP x_O with $S \subset O$ such that an extension of ξ to partition π of O can have positive $\mathrm{mi}_\pi(x_O)$.

4.3 Complete local integration

Complete local integration (CLI) is an important concept in this thesis as positive CLI will form the criterion distinguishing arbitrary STPs from *entities* in multivariate Markov chains (see Section 5.3).

Definition 53 ((Complete) local integration)**.** *Given a Bayesian network $\{X_i\}_{i\in V}$ and an STP x_O of this network the complete local integration $\iota(x_O)$ of x_O is the minimum SLI over the non-unit partitions $\pi \in \mathfrak{L}(O)\setminus\mathbf{1}_O$:*

$$\iota(x_O) := \min_{\pi\in\mathfrak{L}(O)\setminus\mathbf{1}_O} \mathrm{mi}_\pi(x_O). \tag{4.72}$$

We call an STP x_O completely locally integrated if $\iota(x_O) > 0$.

Remarks:

- The reason for excluding the unit partition $\mathbf{1}_O$ of $\mathfrak{L}(O)$ (where $\mathbf{1}_O = \{O\}$ see Definition 4) is that with respect to it every STP has $\mathrm{mi}_{\mathbf{1}_O}(x_O) = 0$.
- The CLI is the SLI of x_O with respect to the partition with respect to which x_O is least integrated. Maybe more clearly, it is the SLI of x_O with respect to the partition that disintegrates x_O the most. The same idea is also employed by Tononi and Sporns (2003); Tononi (2004); Balduzzi and Tononi (2008) in a non-local setting. It is known as the *weakest link approach* (Ay, 2015) to dealing with multiple levels of integration.

We note here that this is not the only approach that is being discussed. Another approach is to look at weighted averages of all integrations. For a further discussion of this point in the case of non-local integration (or complexity which, on a global level, may well be the same thing) see Ay (2015) and references therein. A full analysis of which approach is best suited for the local integration measure presented here is beyond the scope of this thesis.

4.4 Disintegration

In this section we define the disintegration hierarchy and its refinement-free version. We then prove the disintegration theorem which is the main formal result of this thesis. It exposes a connection between partitions minimising the SLI of a trajectory and the CLI of the blocks of such partitions. More precisely for a given trajectory the blocks of the finest partitions among those leading to a particular value of SLI consist only of completely locally integrated blocks. Conversely *each* completely locally integrated STP is a block in such a finest partition among those leading to a particular value of SLI. The theorem therefore reveals the special role of STPs with positive CLI with respect to an entire trajectory of the system. For our purposes this theorem allows further interpretations of the measure of CLI which will be discussed in Section 5.3.5. We believe however that it will also be of general interest in the study of complex systems fore example due to the relation of SLI and CLI to measures of complexity like multi-information and local information dynamics (Section 2.1).

Definition 54 (Disintegration hierarchy)**.** *Given a Bayesian network $\{X_i\}_{i\in V}$ and a trajectory $x_V \in \mathcal{X}_V$, the disintegration hierarchy of x_V is the set $\mathfrak{D}(x_V) = \{\mathfrak{D}_1, \mathfrak{D}_2, \mathfrak{D}_3, ...\}$ of sets of partitions of x_V with:*

(i)

$$\mathfrak{D}_1(x_V) := \operatorname*{arg\,min}_{\pi\in\mathfrak{L}(V)} \mathrm{mi}_\pi(x_V) \tag{4.73}$$

(ii) and for $i > 1$:

$$\mathfrak{D}_i(x_V) := \operatorname*{arg\,min}_{\pi\in\mathfrak{L}(V)\setminus\mathfrak{D}_{<i}(x_V)} \mathrm{mi}_\pi(x_V). \tag{4.74}$$

where $\mathfrak{D}_{<i}(x_V) := \bigcup_{j<i} \mathfrak{D}_j(x_V)$. We call $\mathfrak{D}_i(x_V)$ the i-th disintegration level.

Remark:

- Note that arg min returns all partitions that achieve the minimum SLI.
- Since the Bayesian networks we use are finite, the partition lattice $\mathfrak{L}(V)$ is finite, the set of attained SLI values is finite, and the number $|\mathfrak{D}|$ of disintegration levels is finite.
- In most cases the Bayesian network contains some symmetries among their mechanisms which cause multiple partitions to attain the same SLI value.

- For each trajectory x_V the disintegration hierarchy $\mathfrak{D}$ then partitions the elements of $\mathfrak{L}(V)$ into subsets $\mathfrak{D}_i(x_V)$ of equal SLI. The levels of the hierarchy have increasing SLI.

Definition 55. *Let $\mathfrak{L}(V)$ be the lattice of partitions of set V and let $\mathfrak{E}$ be a subset of $\mathfrak{L}(V)$. Then for every element $\pi \in \mathfrak{L}(V)$ we can define the set*

$$\mathfrak{E}_{\lhd\pi} := \{\xi \in \mathfrak{E} : \xi \lhd \pi\}. \tag{4.75}$$

That is $\mathfrak{E}_{\lhd\pi}$ is the set of partitions in $\mathfrak{E}$ that are refinements of π.

Definition 56 (Refinement-free disintegration hierarchy)**.** *Given a Bayesian network $\{X_i\}_{i\in V}$, a trajectory $x_V \in \mathcal{X}_V$, and its disintegration hierarchy $\mathfrak{D}(x_V)$ the refinement-free disintegration hierarchy of x_V is the set $\mathfrak{D}^{\blacktriangleleft}(x_V) = \{\mathfrak{D}_1^{\blacktriangleleft}, \mathfrak{D}_2^{\blacktriangleleft}, \mathfrak{D}_3^{\blacktriangleleft}, ...\}$ of sets of partitions of x_V with:*

(i)

$$\mathfrak{D}_1^{\blacktriangleleft}(x_V) := \{\pi \in \mathfrak{D}_1(x_V) : \mathfrak{D}_1(x_V)_{\lhd\pi} = \varnothing\}, \tag{4.76}$$

(ii) and for $i > 1$:

$$\mathfrak{D}_i^{\blacktriangleleft}(x_V) := \{\pi \in \mathfrak{D}_i(x_V) : \mathfrak{D}_{<i}(x_V)_{\lhd\pi} = \varnothing\} \tag{4.77}$$

Remark:

- Each level $\mathfrak{D}_i^{\blacktriangleleft}(x_V)$ in the refinement-free disintegration hierarchy $\mathfrak{D}^{\blacktriangleleft}(x_V)$ consists only of those partitions that neither have refinements at their own nor at any of the preceding levels. So each partition that occurs in the refinement-free disintegration hierarchy at the i-th level is a finest partition that achieves such a low level of SLI or such a high level of disintegration.
- As we will see below, the blocks of the partitions in the refinement-free disintegration hierarchy are the main reason for defining the refinement-free disintegration hierarchy.

Theorem 22 (Disintegration theorem)**.** *Let $\{X_i\}_{i\in V}$ be a Bayesian network, $x_V \in \mathcal{X}_V$ one of its trajectories, and $\mathfrak{D}^{\blacktriangleleft}(x_V)$ the associated refinement-free disintegration hierarchy.*

(i) Then for every $\mathfrak{D}_i^{\blacktriangleleft}(x_V) \in \mathfrak{D}^{\blacktriangleleft}(x_V)$ we find for every $b \in \pi$ with $\pi \in \mathfrak{D}_i^{\blacktriangleleft}(x_V)$ that there are only the following possibilities:

1. *b is a singleton, i.e. $b = \{i\}$ for some $i \in V$, or*
2. *x_b is completely locally integrated, i.e. $\iota(x_b) > 0$.*

(ii) Conversely, for any completely locally integrated STP x_A, there is a partition $\pi^A \in \mathfrak{L}(V)$ and a level $\mathfrak{D}_{i^A}^{\blacktriangleleft}(x_V) \in \mathfrak{D}^{\blacktriangleleft}(x_V)$ such that $A \in \pi^A$ and $\pi^A \in \mathfrak{D}_{i^A}^{\blacktriangleleft}(x_V)$.

Proof. **ad (i)** We prove the theorem by contradiction. For this assume that there is block b in a partition $\pi \in \mathfrak{D}_i^{\blacktriangleleft}(x_V)$ which is neither a singleton nor completely integrated. Let $\pi \in \mathfrak{D}_i^{\blacktriangleleft}(x_V)$ and $b \in \pi$. Assume b is not

a singleton i.e. there exist $i \neq j \in V$ such that $i \in b$ and $j \in b$. Also assume that b is not completely integrated i.e. there exists a partition ξ of b with $\xi \neq \mathbf{1}_b$ such that $\mathrm{mi}_\xi(x_b) \leqslant 0$. Note that a singleton cannot be completely locally integrated as it does not allow for a non-unit partition. So together the two assumptions imply $p_b(x_b) \leqslant \prod_{d\in\xi} p_d(x_d)$ with $|\xi| > 1$. But then

$$\mathrm{mi}_\pi(x_V) = \log \frac{p_V(x_V)}{p_b(x_b)\prod_{c\in\pi\setminus b} p_c(x_c)} \tag{4.78}$$

$$\geqslant \log \frac{p_V(x_V)}{\prod_{d\in\xi} p_d(x_d)\prod_{c\in\pi\setminus b} p_c(x_c)} \tag{4.79}$$

We treat the cases of "$>$" and "$=$" separately. First, let

$$\mathrm{mi}_\pi(x_V) = \log \frac{p_V(x_V)}{\prod_{d\in\xi} p_d(x_d)\prod_{c\in\pi\setminus b} p_c(x_c)}. \tag{4.80}$$

Then we can define $\rho := (\pi\setminus b) \cup \xi$ such that

(i) $\mathrm{mi}_\rho(x_V) = \mathrm{mi}_\pi(x_V)$ which implies that $\rho \in \mathfrak{D}_i(x_V)$ because $\pi \in \mathfrak{D}_i(x_V)$, and
(ii) $\rho \lhd \pi$ which contradicts $\pi \in \mathfrak{D}_i^{\blacktriangleleft}(x_V)$.

Second, let

$$\mathrm{mi}_\pi(x_V) > \log \frac{p_V(x_V)}{\prod_{d\in\xi} p_d(x_d)\prod_{c\in\pi\setminus b} p_c(x_c)}. \tag{4.81}$$

Then we can define $\rho := (\pi\setminus b) \cup \xi$ such that

$$\mathrm{mi}_\rho(x_V) < \mathrm{mi}_\pi(x_V), \tag{4.82}$$

which contradicts $\mathrm{mi}_\pi(x_V) \in \mathfrak{D}_i^{\blacktriangleleft}(x_V)$.

ad (ii) Let $\pi^A := \{A\} \cup \{\{j\}\}_{j\in V\setminus A}$. Since π^A is a partition of V it is an element of some disintegration level $\mathfrak{D}_{i^A}$. Then partition π^A is also an element of the refinement free disintegration level $\mathfrak{D}_{i^A}^{\blacktriangleleft}(x_V)$ as we will see in the following. This is because any refinements must (by construction of π^A break up A into further blocks which means that the local specific integration of all such partitions is higher. Then they must be at lower disintegration level $\mathfrak{D}_k(x_V)$ with $k \geqslant i^A$. Therefore π^A has no refinement at its own or a higher disintegration level. More formally, let $\xi \in \mathfrak{L}(V), \xi \neq \pi^A$ and $\xi \lhd \pi^A$ since π^A only contains singletons apart from A the partition ξ must split the block A into multiple blocks $c \in \xi|_A$. Since $\iota(x_A) > 0$ we know that

$$\mathrm{mi}_{\xi|_A}(x_A) = \log \frac{p_A(x_A)}{\prod_{c\in\xi|_A} p_c(x_c)} > 0 \tag{4.83}$$

so that $\prod_{c\in\xi|_A} p_c(x_c) < p_A(x_A)$ and

$$\mathrm{mi}_\xi(x_V) = \log \frac{p_V(x_V)}{\prod_{c\in\xi|_A} p_c(x_c) \prod_{i\in V\setminus A} p_i(x_i)} \tag{4.84}$$

$$> \log \frac{p_V(x_V)}{p_A(x_A) \prod_{i\in V\setminus A} p_i(x_i)} \tag{4.85}$$

$$= \mathrm{mi}_{\pi^A}(x_V). \tag{4.86}$$

Therefore ξ is on a disintegration level $\mathfrak{D}_k(x_V)$ with $k > i^A$, but this is true for any refinement of π^A so $\mathfrak{D}_{<i^A}(x_V)_{\lhd\pi^A} = \varnothing$ and $\pi^A \in \mathfrak{D}^{\blacktriangleleft}_{i^A}(x_V)$. □

4.5 Symmetries and STPs

In this section we present the behaviour of SLI under permutations of the nodes in the Bayesian network. The behaviour of SLI under such operations can be used to explain the appearance of identical disconnected components on the same disintegration levels in the disintegration hierarchy. We we will see this for simple example systems in Chapter 6. The behaviour under transformations like the permutations is also and important property of formal objects in general and can serve as the starting point for further investigations. In order to be able to express the behaviour of SLI under permutations we first have to define the behaviour of STPs and their probabilities under symmetry operations.

We first define terminology for dealing with subgroups of the symmetric group which is the group of all permutations of a set. In particular we will often restrict the permutations to those that only permute nodes within a subset of the Bayesian network. This is relevant for example if we are dealing with a driven multivariate Markov chain where the driven random variables may be permuted freely among each other but not with the driving random variables.

Then we define the group actions of such permutations on STPs, partitions, and probabilities of STPs and show that they are indeed group actions[3]. We then state clearly what we mean by symmetries of STPs, partitions, and probabilities of STPs. After two helper theorems we finally come to the SLI symmetry theorems (Theorems 30 and 31). The first establishes the behaviour of SLI of STPs x_A under permutations that are symmetries of the probability distribution over X_A i.e. over the nodes that are occupied by the STP. The second then establishes the conditions under which the SLI stays invariant under such permutations. These conditions will be used in our example in Chapter 6. We also anticipate that they can be used to establish further theorems about SLI and CLI in particularly symmetric systems. However this is beyond the scope of this thesis.

[3]Group actions have nothing to do with the actions of agents that are important in the conceptual part of this thesis.

4.5.1 Symmetric group terminology

Definition 57. *Let V be a finite set.*

(i) *A permutation of V is a bijective function $g : V \to V$.*

(ii) *The set of all permutations together with function composition $(g_1 \circ g_2)(i) := g_1(g_2(i))$, form a group called the symmetric group Σ_V.*

(iii) *A subgroup of Σ_V is any subset $\mathfrak{G} \subseteq \Sigma_V$ such that for all $g_1, g_2 \in \mathfrak{G}$ we have $g_1 \circ g_2 \in \mathfrak{G}$ and for every $g \in \mathfrak{G}$, $g^{-1} \in \mathfrak{G}$.*

(iv) *Given a subset $A \subseteq V$ and permutation $g \in \Sigma_V$ define:*

$$g(A) := \{g(i) : i \in A\}. \tag{4.87}$$

(v) *Given a subgroup $\mathfrak{G} \subseteq \Sigma_V$ and an element $i \in V$ define the orbit of i under $\mathfrak{G}$ as the set:*

$$\mathfrak{G}(i) := \{g(i) : g \in \mathfrak{G}\}. \tag{4.88}$$

(vi) *Given a subgroup $\mathfrak{G} \subseteq \Sigma_V$ and a subset $A \subseteq V$ define*

$$\mathfrak{G}(A) := \{g(i) : g \in \mathfrak{G}, i \in A\}. \tag{4.89}$$

(vii) *A subset $A \subseteq V$ is an invariant subset or invariant under the action of group $\mathfrak{G}$ if*

$$\mathfrak{G}(A) = A. \tag{4.90}$$

(viii) *A subset $A \in V$ is a fixed subset or fixed under the action of group $\mathfrak{G}$ if for every $g \in \mathfrak{G}$ and $i \in A$*

$$g(i) = i. \tag{4.91}$$

(ix) *For any a subset $A \in V$ let $\Sigma_A \subseteq \Sigma_V$ denote the subgroup of permutations such that $V \backslash A$ is a fixed subset of Σ_A. I.e. for $g \in \Sigma_A$ we have $g(i) = i$ for all $i \in V \backslash A$.*

(x) *Let V be a Cartesian product $V = J \times T$ and let $\mathfrak{G}_1$ be a subgroup of Σ_J and $\mathfrak{G}_2$ a subgroup of Σ_T. Then we can form the group $\mathfrak{G}_1 \times \mathfrak{G}_2$ which is a subgroup of Σ_V by defining for any $g_1 \in \mathfrak{G}_1, g_1 \in \mathfrak{G}_2$ that $(g_1, g_2)(j, t) = (g_1(j), g_2(t))$.*

Remark:

- Note that for all permutations $g : V \to V$ we have $g(V) = V$, but in general for $A \subset V$ we may have either $g(A) \neq A$ or $g(A) = A$.
- Every subgroup contains the identity $\mathfrak{i}$ of Σ_V.

4.5.2 Actions of the symmetric group on patterns, partitions, and probabilities

Definition 58. *Given a Bayesian network $\{X_i\}_{i\in V}$, a subset $A \in V$ with $\mathcal{X}_i = \mathcal{X}_j$ for all $i, j \in A$, permutations $g, h \in \Sigma_A$, and a pattern $x_A \in \mathcal{X}_A \triangleq \{X_i = x_i\}_{i\in A}$ define the following.*

(i) For individual $i \in A$

$$(X_i = x_i)^g := (X_i = x_{g(i)}) \tag{4.92}$$

we also write x_i^g where there is no danger of confusion.

(ii) Furthermore

$$(X_i = x_{g(i)})^h := (X_i = x_{g(h(i))}) \tag{4.93}$$

we also write $(x^g)_i^h$ where there is no danger of confusion.

(iii) For $B \subseteq A$ the STP x_A

$$\{X_i = x_i\}_{i\in B}^g := \{(X_i = x_i)^g\}_{i\in B} \tag{4.94}$$

we also write x_B^g where there is no danger of confusion.

Remarks:

- Note that we require the state spaces of all the random variables in A to be equal in order for x_i^g to be well defined for all $i \in A$ and all $x_i \in \mathcal{X}_i$.
- In words the pattern $\{X_i = x_i\}_{i\in B}^g$ fixes the random variables at $i \in B$ to the values that x_A defines at $g(i)$. Since the state spaces are identical by assumption this is well defined.
- These definitions are based on the full notation of patterns because the shorthand notation does not afford the necessary expressiveness. For calculations in the rest of the section we will often resort to the full notation but at the same time try to use the visually less demanding shorthand where possible.
- A simple example: let $V = A = \{1, 2\}$ and $g(1) = 2, g(2) = 1$ then:

$$\begin{aligned} \{X_1 = x_1, X_2 = \bar{x}_2\}^g &= \{(X_1 = x_1)^g, (X_2 = \bar{x}_2)^g\} && (4.95)\\ &= \{X_1 = \bar{x}_{g(1)}, X_2 = x_{g(2)}\} && (4.96)\\ &= \{X_1 = \bar{x}_2, X_2 = x_1\}. && (4.97) \end{aligned}$$

We used the bar over $\bar{x}_2$ to highlight the movement of the values. This will be useful when we look at marginalisations later. Note that $(X_1 = x_1)^g$ does not reflect the value that $g(1)$ maps to (there is no bar over the x) but $X_1 = \bar{x}_{g(1)}$ does. This means that for the latter notation we already have to know the result of $g(1)$ in order to know whether to put a bar over x or not.

- These definition provide a slight adaptation of the definition of "x^g" in Ceccherini-Silberstein and Coornaert (2009). The reason for the counter intuitive rule in 4.93 also stems from the close relation of the present definition to the idea behind the definition in that publication. In their case x is a function acting on the indices and x^g is defined as the function taking as argument $g(i)$. Consequently, the function $(x^g)^h$ takes as argument $g(h(i))$. In our case x is not a function (and cannot directly be made into one as the state spaces of the random variables in our Bayesian network may differ unlike in Ceccherini-Silberstein and Coornaert (2009)) so that we emulate similar behaviour with the above rule.

Definition 59 (Action on a STP)**.** *Given a Bayesian network $\{X_i\}_{i\in V}$ a subset $A \in V$ with $\mathcal{X}_i = \mathcal{X}_j$ for all $i, j \in A$, a permutation $g \in \Sigma_A$, a pattern x_A, and a subset $C \in A$ define the action of g on x_C, as: In full notation:*

$$\tilde{g}\{X_i = x_i\}_{i\in C} := \{X_i = x_{g^{-1}(i)}\}_{i\in C}. \tag{4.98}$$

In short notation:

$$\tilde{g}x_C := x_C^{g^{-1}}. \tag{4.99}$$

If not necessary we write just g instead of $\tilde{g}$.

Remark:

- We use $\tilde{g}$ here to indicate that this is another object than the permutation g which is a function on index sets. For the subsequent proof that the above definition is an action this distinction is necessary. Beyond this proof it is always clear that it is another object from the context.
- The result of the action of g on x_C is then another pattern $\bar{x}_C$ where the new value $\bar{x}_i$ of the random variable X_i at node i is now the value $x_j = x_{g^{-1}(i)}]$ originally fixed for the random variable X_j at node $j = g^{-1}(i)$. One can think of this construction in analogy to "shifting a function $f : \mathcal{X} \to \mathcal{Y}$ to the right" by a constant d by defining $\bar{f}(x) := f(x - d)$. In order to get the function to move in the positive direction by d we its negative $-d$ to the argument. This has the desired effect. Similarly to transform the STP by g we act on the indices with its inverse g^{-1}. This is common practice in defining group actions, for a similar construction see Ceccherini-Silberstein and Coornaert (2009).
- It might be redundant to define the action of gx_C on top of the previous definition of x_C^g which is an equivalent construction. The detour is presented here because it allows for the use of the standard construction of the action on probabilities in Definition 61.

Theorem 23. *Given a Bayesian network $\{X_i\}_{i\in V}$ a subset $A \in V$ with $\mathcal{X}_i = \mathcal{X}_j$ for all $i, j \in A$, the action on STPs of Definition 59 is a group action of the group Σ_A on the set of all STP $\bigcup_{B\subseteq A} \mathcal{X}_B$. This means that for all $g, h \in \Sigma_A$ and all $x_C \in \bigcup_{B\subseteq A} \mathcal{X}_B$ we have*

(i)

$$\tilde{g}x_C \in \bigcup_{B\subseteq A} \mathcal{X}_B, \tag{4.100}$$

(ii)

$$\tilde{h}(\tilde{g}x_C) = \widetilde{(h \circ g)}x_C, \tag{4.101}$$

(iii)

$$\tilde{\mathrm{i}}x_C = x_C. \tag{4.102}$$

Proof. **ad (i)** Note that $\tilde{g}x_C = \{X_i = x_{g^{-1}(i)}\}_{i\in C}$. Recall $\mathcal{X}_i = \mathcal{X}_j$ for all $i, j \in A$ and $g \in \Sigma_A$ such that $g^{-1} \in \Sigma_A$. Then we know that for all $i \in C$ $x_{g^{-1}(i)} \in \mathcal{X}_{g^{-1}(i)} = \mathcal{X}_i$ thus $\bar{x}_C$ with $\bar{x}_i = x_{g^{-1}(i)}$ is a pattern $\bar{x}_C \in \mathcal{X}_C \subseteq \bigcup_{B\subseteq A} \mathcal{X}_B$.

ad (ii) In the full notation we have:

$$\begin{aligned}
\tilde{h}(\tilde{g}x_C) &= \tilde{h}\{X_i = x_i\}_{i\in C}^{g^{-1}} && (4.103)\\
&= \tilde{h}\{(X_i = x_i)^{g^{-1}}\}_{i\in C} && (4.104)\\
&= \{X_i = x_{g^{-1}(i)}\}_{i\in C}^{h^{-1}} && (4.105)\\
&= \{(X_i = x_{g^{-1}(i)})^{h^{-1}}\}_{i\in C} && (4.106)\\
&= \{(X_i = x_{(g^{-1}(h^{-1})(i)})\}_{i\in C} && (4.107)\\
&= \{(X_i = x_{(g^{-1}\circ h^{-1})(i)})\}_{i\in C} && (4.108)\\
&= \{(X_i = x_{(h\circ g)^{-1}(i)})\}_{i\in C} && (4.109)\\
&= \{(X_i = x_i)\}_{i\in C}^{(h\circ g)^{-1}} && (4.110)\\
&= \widetilde{(h \circ g)}\{(X_i = x_i)\}_{i\in C} && (4.111)\\
&= \widetilde{(h \circ g)}x_C. && (4.112)
\end{aligned}$$

ad (iii) Note $\tilde{\mathrm{i}}x_C = x_C^{\mathrm{i}} = \{X_i = x_{\mathrm{i}(i)}\}_{i\in C} = x_C$.

□

Definition 60 (Action on a partition)**.** *Given a set V, a subset $A \subseteq V$, a partition $\pi \in \mathfrak{L}(A)$, and a permutation $g \in \Sigma_A$ define the action of g on π by*

$$\hat{g}\pi := \{g(b) \subseteq V : b \in \pi\}. \tag{4.113}$$

Again if it is not necessary we just write $g\pi$ instead of $\hat{g}\pi$.

Theorem 24. *Given a set V, a subset $A \subseteq V$ a partition $\pi \in \mathfrak{L}(A)$, and a permutation $g \in \Sigma_A$, the action on partitions of Definition 60 is a group action of Σ_A on the set of all partitions $\pi \in \mathfrak{L}(A)$. This means that for all $g, h \in \Sigma_A$ and all $\pi \in \mathfrak{L}(A)$ we have*

(i)

$$\hat{g}\pi \in \mathfrak{L}(A) \tag{4.114}$$

(ii)

$$\hat{h}(\hat{g}\pi) = \widehat{(h \circ g)}\pi \tag{4.115}$$

(iii)

$$\mathrm{i}\pi = \pi \tag{4.116}$$

Proof. **ad (i)** Note that since $g \in \Sigma_A$ we have $g(i) \in A$ for every $i \in A$. Therefore $\hat{g}\pi = \{g(b) \subseteq V : b \in \pi\} = \{g(b) \subseteq A : b \in \pi\}$, so all blocks of π are mapped to subsets of A.

To show that $\hat{g}\pi$ is a partition we need to show two things. First, that for all $b_1, b_2 \in \pi$ we have $g(b_1) \cap g(b_2) = \varnothing$. Note $b_1 \cap b_2 = \varnothing$ by assumption and $g : V \to V$ is injective (even bijective) so that we have for all $i, j \in V$, $g(i) \neq g(j)$ if $i \neq j$. Then for all $i \in b_1, j \in b_2$ always $i \neq j$ and $g(i) \neq g(j)$ so $g(b1) \cap g(b_2) = \varnothing$.

Second, we have to show that $\bigcup_{c \in \hat{g}\pi} c = A$ which follows from surjectivity of g. Pick any $j \in A$ then there exists $i \in A$ such that $g(i) = j$. Since π is a partition of A, for each such $i \in A$ there is a block $b \in \pi$ with $i \in \pi$ such that $j \in \hat{g}\pi$.

ad (ii) Note:

$$\begin{aligned}
\hat{h}(\hat{g}\pi) &= \hat{h}\{g(b) \subseteq A : b \in \pi\} && (4.117)\\
&= \{h(g(b)) \subseteq A : b \in \pi\} && (4.118)\\
&= \{(h \circ g)(b) \subseteq A : b \in \pi\} && (4.119)\\
&= \widehat{(h \circ g)}\pi && (4.120)
\end{aligned}$$

ad (iii) Note $\hat{\mathrm{i}}\pi = \{\mathrm{i}(b) \subseteq A : b \in \pi\} = \pi$.

□

Theorem 25. *Given a set V, a subset $A \subseteq V$, partitions $\pi, \xi \in \mathfrak{L}(A)$, and a permutation $g \in \Sigma_A$ the action of g on the partitions preserves the refinement relation. Formally:*

$$\pi \preccurlyeq \xi \Leftrightarrow g\pi \preccurlyeq g\xi. \tag{4.121}$$

Proof. From $i \equiv_\pi j$ which just means that there exists $b in \pi$ with $i, j \in b$ we get $g(i) \equiv_{g\pi} g(j)$ since membership of sets is preserved i.e. if $i \in b$ then $g(i) \in g(b)$. So if $i \equiv_\pi j \Rightarrow i \equiv_\xi j$ then also $g(i) \equiv_{g\pi} g(j) \Rightarrow g(i) \equiv_{g\xi} g(j)$. So by Definition 13 $g\pi \preccurlyeq g\xi$. □

Remark:

- Visually, this means that the Hasse diagram of a set of transformed partitions is the same as that of the non-transformed partitions.

Theorem 26. *Given a set V, a subset $A \subseteq V$, partitions $\pi \in \mathfrak{L}(A)$, and a permutation $g \in \Sigma_A$ the action of g on the partitions preserves the cardinality $|\pi|$. Formally:*

$$|\pi| = |g\pi|. \tag{4.122}$$

Proof. We have seen that $g\pi$ is a partition as well. By definition it is only composed out of the images of the blocks, so there cannot be more blocks in $g\pi$ than in π. We also know that g is injective so no two elements can be mapped to the same element and by extension no tow blocks can be mapped to the same block. Therefore there cannot be fewer blocks in $g\pi$ than in π. □

Definition 61 (Action on probability distributions)**.** *Given a Bayesian network $\{X_i\}_{i\in V}$ a subset $A \in V$ with $\mathcal{X}_i = \mathcal{X}_j$ for all $i, j \in A$, and any permutation $g \in \Sigma_A$ define the action of g on the probability distribution $p_V : \mathcal{X}_V \to \mathcal{X}_V$ by setting for for each $x_V \in \mathcal{X}_V$:*

$$(\check{g}p_V)(x_V) := \Pr(g^{-1}\{X_i = x_i\}_{i\in V}) \tag{4.123}$$

$$= \Pr(\{X_i = x_{g(i)}\}_{i\in V}). \tag{4.124}$$

If there is no danger of confusion we also write in short notation of the patterns $(\check{g}p_V)(x_V) = p_V(x_V^g)$. Also if not necessary we will write (gp_V) instead of $(\check{g}p_V)$.

Remark:

- So the probability distribution $\check{g}p_V$ resulting from the action of g on p_V assigns x_V the probability that was originally assigned to $g^{-1}x_V$. The latter is the trajectory x_V^g that we obtain by moving the values $x_{g(i)}$ to nodes i.
- The action of g on p_V as defined here corresponds to the usual way group actions are defined on functions. Namely by making the inverse g^{-1} act on the argument to the function.
- We only require that the state spaces of the random variables $\{X_i\}_{i\in A}$ whose indices g does not keep fixed are equal. This allows us to deal with situations where some random variables have different state spaces or do not exhibit the same symmetric structure as those in A. This can be useful in the case of driven Markov chains.

Theorem 27. *Given a Bayesian network $\{X_i\}_{i\in V}$ a subset $A \in V$ with $\mathcal{X}_i = \mathcal{X}_j$ for all $i, j \in A$, and a permutation $g \in \Sigma_A$, the action on probability distributions of Definition 61 is a group action of Σ_A on the set of all probability distributions $p_V \in \mathcal{P}(\mathcal{X}_V)$. This means that for all $g, h \in \Sigma_A$ and all p_V we have*

(i)

$$\check{g}p_V \in \mathcal{P}(\mathcal{X}_V) \tag{4.125}$$

(ii)

$$\check{h}(\check{g}p_V) = \widecheck{(h \circ g)}p_V \tag{4.126}$$

(iii)

$$\check{\imath}p_V = p_V \tag{4.127}$$

Proof. **ad (i)** First, note that since $g \in \Sigma_A$ so that $g^{-1} \in \Sigma_A$ and $\mathcal{X}_k = \mathcal{X}_l$ for all $k, l \in A$ we have $x_{g^{-1}(i)} \in \mathcal{X}_i$ for all $i \in A$. Also because of $g^{-1} \in \Sigma_A$ we have $g^{-1}(i) = i$ for $i \in V\backslash A$, so actually $x_{g^{-1}(i)} \in \mathcal{X}_i$ for all $i \in V$. This makes x_V^g a valid pattern.

Second, show that for all $x_V \in \mathcal{X}_V$ we have $(\check{g}p_V)(x_V) \in [0, 1]$. This follows from x_V^g being a valid pattern for each x_V and the probability of any valid pattern being in the interval $[0, 1]$.

Third, we show that the transformed distribution $\check{g}p_V$ is normalised. Since the action of g on x_V realises a bijective function on $\mathcal{X}_V$ (g has an

inverse, and its inverse has an inverse) we can substitute x_V with $\check{g}x_V$ as in the following:[4]

$$\sum_{x_V \in \mathcal{X}_V} (\check{g}p_V)(x_V) = \bigoplus_{j \in A} \sum_{x_j \in \mathcal{X}_j} \Pr\left(g^{-1}\{X_i = x_i\}_{i \in V}\right) \tag{4.128}$$

$$= \bigoplus_{j \in A} \sum_{x_{g^{-1}(j)} \in \mathcal{X}_j} \Pr\left(g^{-1}\{X_i = x_{g^{-1}(i)}\}_{i \in V}\right) \tag{4.129}$$

$$= \bigoplus_{j \in A} \sum_{x_j \in \mathcal{X}_j} \Pr\left(g^{-1}\{X_i = x_{g^{-1}(i)}\}_{i \in V}\right) \tag{4.130}$$

$$= \bigoplus_{j \in A} \sum_{x_j \in \mathcal{X}_j} \Pr\left(g^{-1}(g\{X_i = x_i\}_{i \in V})\right) \tag{4.131}$$

$$= \bigoplus_{j \in A} \sum_{x_j \in \mathcal{X}_j} \Pr\left(\{X_i = x_i\}_{i \in V}\right) \tag{4.132}$$

$$= 1. \tag{4.133}$$

ad (ii) Note, for all x_V:

$$(\check{h}(\check{g}p_V))(x_V) = (\check{g}p_V)(h^{-1}x_V) \tag{4.134}$$

$$= p_V(g^{-1}(h^{-1}x_V)) \tag{4.135}$$

$$= p((g^{-1} \circ h^{-1})x_V) \tag{4.136}$$

$$= p((h \circ g)^{-1}x_V) \tag{4.137}$$

$$= (\widetilde{(h \circ g)}p_V)(x_V). \tag{4.138}$$

ad (iii) For all $x_V \in \mathcal{X}_V$, we have $\check{\imath}p_V(x_V) = p_V(\imath x_V) = p_v(x_V)$.

□

Definition 62 (Symmetries). *Given a Bayesian network $\{X_i\}_{i \in V}$ a subset $A \in V$ with $\mathcal{X}_i = \mathcal{X}_j$ for all $i, j \in A$, a permutation $g \in \Sigma_A$, a pattern x_A, a subset $C \in A$, and a partition π of C.*

(i) We say g is a symmetry of x_C or x_C is invariant under g if

$$gx_C = x_C, \tag{4.139}$$

and call the group of all symmetries of x_C the symmetry group of x_C.

(ii) We say g is a symmetry of π or π is invariant under g if

$$g\pi = \pi, \tag{4.140}$$

and call the group of all symmetries of π the symmetry group of π.

(iii) We say g is a symmetry of $p_C : \mathcal{X}_C \to [0, 1]$ or p_C is invariant under g if

$$gp_C = p_C, \tag{4.141}$$

and call the group of all symmetries of p_C the symmetry group of p_C.

Remark:

- We can only have $gx_C = x_C$ or $g\pi = \pi$ if $gC = C$.

[4] We use $\bigoplus$ here to denote that for each j there is a sum over x_j. This is to keep the notation somewhat short, and clear.

4.5.3 Transformation of SLI for invariant probability distributions

Theorem 28. *Given a Bayesian network $\{X_i\}_{i\in V}$ a subset $A \in V$ with $\mathcal{X}_i = \mathcal{X}_j$ for all $i, j \in A$, a permutation $g \in \Sigma_A$, and a subset $B \subseteq A$ we have for every $x_B \in \mathcal{X}_B$:*

$$(gp_B)(x_B) = \Pr\left(\{X_i = x_{g(i)}\}_{i\in g^{-1}(B)}\right). \tag{4.142}$$

We can also write this as

$$(gp_B)(x_B) = p_{g^{-1}(B)}(x^g_{g^{-1}(B)}). \tag{4.143}$$

Furthermore, if $B = A$ we get:

$$(gp_A)(x_A) = p_A(x^g_A). \tag{4.144}$$

Proof. By definition we have:

$$(gp_B)(x_B) = \bigoplus_{j\in V\setminus B} \sum_{\bar{x}_j\in\mathcal{X}_j} \Pr\left(g^{-1}\left(\{X_i = x_i\}_{i\in B}, \{X_i = \bar{x}_i\}_{i\in V\setminus B}\right)\right) \tag{4.145}$$

$$= \bigoplus_{j\in V\setminus B} \sum_{\bar{x}_j\in\mathcal{X}_j} \Pr\left(\{X_i = x_i\}^g_{i\in B}, \{X_i = \bar{x}_i\}^g_{i\in A\setminus B}, \{X_i = \bar{x}_i\}_{i\in V\setminus A}\right) \tag{4.146}$$

$$= \bigoplus_{j\in A\setminus B} \sum_{\bar{x}_j\in\mathcal{X}_j} \Pr\left(\{X_i = x_i\}^g_{i\in B}, \{X_i = \bar{x}_i\}^g_{i\in A\setminus B}\right) \tag{4.147}$$

$$= \bigoplus_{j\in A\setminus B} \sum_{\bar{x}_j\in\mathcal{X}_j} \Pr\Big(\{X_i = x_i\}^g_{i\in g^{-1}(B)\cap B'}, \{X_i = \bar{x}_i\}^g_{i\in g^{-1}(B)\setminus B'}, \{X_i = \bar{x}_i\}^g_{i\in g^{-1}(A\setminus B)\cap(A\setminus B)'}, \{X_i = x_i\}^g_{i\in g^{-1}(A\setminus B)\setminus(A\setminus B)}\Big) \tag{4.148}$$

$$= \bigoplus_{j\in g^{-1}(A\setminus B)} \sum_{\bar{x}_j\in\mathcal{X}_j} \Pr\Big(\{X_i = x_{g(i)}\}_{i\in g^{-1}(B)\cap B}, \{X_i = x_{g(i)}\}_{i\in g^{-1}(B)\setminus B}, \{X_i = \bar{x}_{g(i)}\}_{i\in g^{-1}(A\setminus B)\cap(A\setminus B)}, \{X_i = \bar{x}_{g(i)}\}_{i\in g^{-1}(A\setminus B)\setminus(A\setminus B)}\Big) \tag{4.149}$$

$$= \bigoplus_{j\in A\setminus B} \sum_{\bar{x}_j\in\mathcal{X}_j} \Pr\left(\{X_i = x_{g(i)}\}_{i\in g^{-1}(B)}, \{X_i = \bar{x}_{g(i)}\}_{i\in g^{-1}(A\setminus B)}\right) \tag{4.150}$$

$$= \Pr\left(\{X_i = x_{g(i)}\}_{i\in g^{-1}(B)}\right) \tag{4.151}$$

$$= \Pr\left(g^{-1}\{X_i = x_i\}_{i\in g^{-1}(B)}\right) \tag{4.152}$$

$$=: p_{g^{-1}(B)}(x^g_{g^{-1}(B)}). \tag{4.153}$$

Note the movement of the bar and with it the change of summation indices from $A\setminus B$ to $g^{-1}(A\setminus B)$.

Since $g \in \Sigma_A$ we have $g^{-1}(A) = A$ such that if $B = A$ it follows immediately that

$$p_{g^{-1}(A)}(x^g_{g^{-1}(A)}) = p_A(x^g_A). \tag{4.154}$$

□

Remark:

- The main advantage of the full notation is the clear separation between the action of g on the index of the assigned values i.e. $x_i \mapsto x_{g(i)}$ and the action on the sets of indices that are affected by a pattern i.e. $B \mapsto g(B)$. In the shorthand notation this separation may easily get lost during manipulations of the equations.
- This theorem shows the effect of the action of a group element $g \in \Sigma_A$ on the (marginalised) probability of a pattern within A. Marginalisation is used in the usual way but now for the transformed probability distribution gp_V instead of for p_V.

Theorem 29. *Given a Bayesian network $\{X_i\}_{i\in V}$, a subset $A \subset V$ of the index set V, and a group of permutations $\mathfrak{G} \subseteq \Sigma_A$. If for all $g \in \mathfrak{G}$ we have:*

$$gp_V = p_V, \tag{4.155}$$

then for all $g \in \mathfrak{G}$ also:

$$gp_A = p_A. \tag{4.156}$$

Proof.

$$(gp_A)(x_A) = \bigoplus_{j\in V\setminus A} \sum_{\bar{x}_j\in\mathcal{X}_j} \Pr\left(g^{-1}\left(\{X_i = x_i\}_{i\in A}, \{X_i = \bar{x}_i\}_{i\in V\setminus A}\right)\right) \tag{4.157}$$

$$= \bigoplus_{j\in V\setminus A} \sum_{\bar{x}_j\in\mathcal{X}_j} \Pr\left(\{X_i = x_i\}^g_{i\in A}, \{X_i = \bar{x}_i\}^g_{i\in V\setminus A}\right) \tag{4.158}$$

$$= \bigoplus_{j\in V\setminus A} \sum_{\bar{x}_j\in\mathcal{X}_j} \Pr\left(\{X_i = x_i\}^g_{i\in A}, \{X_i = \bar{x}_i\}_{i\in V\setminus A}\right) \tag{4.159}$$

$$= \bigoplus_{j\in V\setminus A} \sum_{\bar{x}_j\in\mathcal{X}_j} \Pr\left(\{X_i = x_i\}_{i\in A}, \{X_i = \bar{x}_i\}_{i\in V\setminus A}\right) \tag{4.160}$$

$$= p_A(x_A). \tag{4.161}$$

□

Remark:

- This theorem shows that if we find subgroups of Σ_V affecting only a subset $A \subseteq V$ of the nodes in Bayesian network (i.e. any subgroup of Σ_A) and we have $gp_V = p_V$ i.e. the joint probability distribution over X_V is invariant with respect to the action of this subgroup, then the joint probability distribution p_A over the subset of nodes is also invariant with respect to the action of this group.

Theorem 30 (SLI symmetry theorem)**.** *Given a Bayesian network $\{X_i\}_{i\in V}$, a subset $A \subset V$ of the index set V, and a group of permutations $\mathfrak{G} \subseteq \Sigma_A$ such that for all $g \in \mathfrak{G}$ we have:*

$$gp_A = p_A \tag{4.162}$$

then for all $x_A \in \mathcal{X}_A$ and all $g \in \mathfrak{G}$:

$$\mathrm{mi}_{g\pi}(x_A) = \mathrm{mi}_\pi(x^g_A) \tag{4.163}$$

Proof. Note first:

$$\mathrm{mi}_{g\pi}(x_A) = \log \frac{p_A(x_A)}{\prod_{b\in g\pi} p_b(x_b)} \tag{4.164}$$

$$= \log \frac{p_A(x_A)}{\prod_{c\in\pi} p_{g(c)}(x_{g(c)})}. \tag{4.165}$$

Now look at the block probabilities individually and use the invariance of p_A:

$$p_{g(c)}(x_{g(c)}) = \sum_{\bar{x}_{V\setminus g(c)}} p_V(x_{g(c)}, \bar{x}_{A\setminus g(c)}, \bar{x}_{V\setminus A}) \tag{4.166}$$

$$= \bigoplus_{j\in V\setminus g(c)} \sum_{\bar{x}_j\in\mathcal{X}_j} \Pr\left(\{X_i = x_i\}_{i\in g(c)}, \{X_i = \bar{x}_i\}_{i\in V\setminus g(c)}\right) \tag{4.167}$$

$$= \bigoplus_{j\in V\setminus g(c)} \sum_{\bar{x}_j\in\mathcal{X}_j} \Pr\left(g^{-1}\left(\{X_i = x_i\}_{i\in g(c)}, \{X_i = \bar{x}_i\}_{i\in V\setminus g(c)}\right)\right) \tag{4.168}$$

$$= (gp_{g(c)})(x_{g(c)}) \tag{4.169}$$

$$= Pr\left(\{X_i = x_{g(i)}\}_{i\in g^{-1}(g(c))}\right) \tag{4.170}$$

$$= Pr\left(\{X_i = x_{g(i)}\}_{i\in c}\right) \tag{4.171}$$

$$= p_c(x_c^g). \tag{4.172}$$

Plugging the block probabilities and $p_A(x_A) = p_A(x_A^g)$ into Eq. (4.165) concludes the proof. □

Remark:

- This theorem concerns the reaction of the SLI to a transformation of the partition by a permutation under which the probability distribution is invariant.
- Note that we can substitute $x_A \mapsto x_A^{g^{-1}}$ and get $\mathrm{mi}_\pi(x_A) = \mathrm{mi}_{g\pi}(gx_A)$.

Theorem 31 (SLI symmetry corollary)**.** *Under the assumptions of Theorem 30, let $\mathfrak{H}$ be a subgroup of $\mathfrak{G}$.*

(i) If for some x_A we have for all $h \in \mathfrak{H}$:

$$x_A^h = x_A \tag{4.173}$$

then for all $h \in \mathfrak{H}$:

$$\mathrm{mi}_{h\pi}(x_A) = \mathrm{mi}_\pi(x_A). \tag{4.174}$$

(ii) If $x_A^h \neq x_A$ but for all $b \in \pi$ we have

$$p_b(x_b^h) = p_b(x_b) \tag{4.175}$$

then for all $g \in \mathfrak{G}$:

$$\mathrm{mi}_{g\pi}(x_A) = \mathrm{mi}_\pi(x_A). \tag{4.176}$$

(iii) If $x_A^g \neq x_A$ and there exists a $b \in \pi$ with $p_b(x_b^g) \neq p_b(x_b)$, but for all $g \in \mathfrak{G}$ we have

$$\prod_{b\in\pi} p_b(x_b^g) = \prod_{b\in\pi} p_b(x_b) \tag{4.177}$$

then for all $g \in \mathfrak{G}$:

$$\mathrm{mi}_{g\pi}(x_A) = \mathrm{mi}_\pi(x_A). \tag{4.178}$$

Proof. **ad (i):** Follows directly from Theorem 30 by plugging in $x_A^h = x_A$ which also implies $x_b^h = x_b$ for all $b \in \pi$.

ad (ii): Note that $p(x_A^h) = p(x_A)$ since $hp_A = p_A$ and $(hp_A)(x_A) = p_A(x_A^h)$ according to Theorem 28. Then:

$$\mathrm{mi}_{h\pi}(x_A) = \log \frac{p_A(x_A^h)}{\prod_{b\in\pi} p_b(x_b^h)} = \log \frac{p_A(x_A)}{\prod_{b\in\pi} p_b(x_b)}. \tag{4.179}$$

Where we used that $p_b(x_b^h) = p_b(x_b)$ for all $b \in \pi$ by assumption.

ad (iii): Just like for (ii) we have

$$\mathrm{mi}_{h\pi}(x_A) = \log \frac{p_A(x_A)}{\prod_{b\in\pi} p_b(x_b^h)} = \log \frac{p_A(x_A)}{\prod_{b\in\pi} p_b(x_b)}, \tag{4.180}$$

where, this time we used that $\prod_{b\in\pi} p_b(x_b^h) = \prod_{b\in\pi} p_b(x_b)$ by assumption.

□

Remarks:

- $\mathfrak{H}$ needs to be a subgroup of $\mathfrak{G}$ because else we don't have $p_{h(b)}(x_{h(b)}) = p_b(x_b^h)$. Recall that this only holds for elements $g \in \mathfrak{G}$ because for those $gp_A = p_A$ and only in that case we have Eq. (4.172).
- These equalities all concern the consequences of Theorem 30 for partitions of a single STP x_A. The consequences across differing trajectories are not covered.
- The three statements can be seen to describe three levels of conditions which imply the equality of SLI. In all three cases, due to the fact that $\mathfrak{G}$ leaves $p_A(x_A)$ invariant we have $p_{gb}(x_{gb}) = p_b(x_b^g)$ for all $b \in \pi$. In the first case and on what could be called the lowest level if a subgroup of $\mathfrak{G}$ also leaves x_A invariant then $x_b^h = x_b$ and therefore $p_b(x_b^h) = p_b(x_b)$ for all $b \in \pi$.

  The second states that if there is no invariance of x_A then for some other reason the probability distribution p_A may still be such that all block probabilities are invariant i.e. $p_b(x_b^h) = p_b(x_b)$ for all $b \in \pi$ and we still get equal SLI.

  The third then shows that if there is no invariance of x_A or of all the p_b then it may still be the case that the product $\prod_{b\in\pi} p_b(x_b^h) = \prod_{b\in\pi} p_b(x_b)$ being equal even if the individual terms differ.

4.6 Symmetries and Markov chains

Here, we state and prove in our own notation three theorems on symmetries and Markov chains. We first define what we mean by a symmetry of a Markov matrix. These are purely spatial permutations that commute with the Markov matrix. We then show the consequence for the individual entries of the Markov matrix of such a symmetry. Then we show that a symmetry of the initial distribution which is also a symmetry of the Markov matrix becomes a symmetry of the joint probability distribution of the entire Bayesian network i.e. of the probability distribution over the trajectories. This theorem is well known and only presented for quick reference since we use it in Section 6.3.6. Finally, we look at the special case where only some of the spatial indices are permuted and (for example) the driving variables are left alone. In that case, as long as the interactions with the driving variables obey the symmetry as well (in the sense established by the theorem) the symmetry of an initial distribution is also extended to the joint probability distribution over the entire network. This theorem is provided as a connection to simulations of driven systems which exhibit life-like behaviour. If the drive obeys the condition presented here the SLI symmetry theorems can also be used for such systems and not only for cellular automata. We give a short example of a thermostat-like system in the remark. For the rest of the thesis, this theorem is not important.

Definition 63 (Spatial symmetries of Markov matrices)**.** *Let $\{X_i\}_{i\in V}$ be a multivariate Markov chain with $V = J \times T$ and let $g \in \Sigma_J \times \{\mathfrak{i}\}$.*[5] *Then we say g is a symmetry of P_{t+1} or P_{t+1} is invariant under g if for all probability distributions $p_{V_t} : \mathcal{X}_{V_t} \to [0,1]$ we have*

$$gp_{V_{t+1}} = g(P_{t+1}p_{V_t}) = P_{t+1}(gp_{V_t}), \tag{4.181}$$

and call the group of all symmetries of P_{t+1} the symmetry group of P_{t+1}.

Theorem 32. *Let $\{X_i\}_{i\in V}$ be a multivariate Markov chain with $V = J \times T$.*

(i) Let $\mathfrak{G}$ be a subgroup of the symmetry group of P_{t+1}. Then for every $g \in \mathfrak{G}$:

$$p_{V_{t+1}}(x^g_{V_{t+1}}|x^g_{V_t}) = p_{V_{t+1}}(x_{V_{t+1}}|x_{V_t}) \tag{4.182}$$

which is equivalent to

$$p_{V_{t+1}}(x^g_{V_{t+1}}|\hat{x}_{V_t}) = p_{V_{t+1}}(x_{V_{t+1}}|\hat{x}^{g^{-1}}_{V_t}). \tag{4.183}$$

(ii) Conversely, if for all elements g of a subgroup $\mathfrak{G}$ of $\Sigma_J \times \{\mathfrak{i}\}$ we have

$$p_{V_{t+1}}(x^g_{V_{t+1}}|\hat{x}_{V_t}) = p_{V_{t+1}}(x_{V_{t+1}}|\hat{x}^{g^{-1}}_{V_t}). \tag{4.184}$$

then $\mathfrak{G}$ is a symmetry group of P_{t+1}.

[5] Recall that $\mathfrak{i}$ indicates the identity element of a group.

Proof. **ad (i):** By assumption Eq. (4.181) holds for any choice of p_{V_t} so we choose for arbitrary $\hat{x}_{V_t} \in \mathcal{X}_{A_t}$

$$p_{V_t}(x_{V_t}) = \delta_{\hat{x}_{V_t}}(x_{V_t}), \tag{4.185}$$

and plug this first into the left hand side of Eq. (4.181):

$$\begin{aligned}
(gp_{V_{t+1}})(x_{V_{t+1}}) &= p_{V_{t+1}}(x^g_{V_{t+1}}) && (4.186)\\
&= \sum_{\bar{x}_{V_t}} p_{V_{t+1}}(x^g_{V_{t+1}}|\bar{x}_{V_t}) p_{V_t}(\bar{x}_{V_t}) && (4.187)\\
&= \sum_{\bar{x}_{V_t}} p_{V_{t+1}}(x^g_{V_{t+1}}|\bar{x}_{V_t}) \delta_{\hat{x}_{V_t}}(\bar{x}_{V_t}) && (4.188)\\
&= p_{V_{t+1}}(x^g_{V_{t+1}}|\hat{x}_{V_t}). && (4.189)
\end{aligned}$$

Then into the right hand side of Eq. (4.181):

$$\begin{aligned}
(P_{t+1}(gp_{V_t}))(x_{V_{t+1}}) &= \sum_{\bar{x}_{V_t}} p_{V_{t+1}}(x_{V_{t+1}}|\bar{x}_{V_t}) p_{V_t}(\bar{x}^g_{V_t}) && (4.190)\\
&= \sum_{\bar{x}_{V_t}} p_{V_{t+1}}(x_{V_{t+1}}|\bar{x}_{V_t}) \delta_{\hat{x}_{V_t}}(\bar{x}^g_{V_t}) && (4.191)\\
&= p_{V_{t+1}}(x_{V_{t+1}}|\hat{x}^{g^{-1}}_{V_t}), && (4.192)
\end{aligned}$$

where we used that $\hat{x}_{V_t} = \bar{x}^g_{V_t} \Leftrightarrow \hat{x}^{g^{-1}}_{V_t} = \bar{x}_{V_t}$. So we have for all $x_{V_{t+1}}$ and all $\hat{x}_{V_t}$:

$$p_{V_{t+1}}(x^g_{V_{t+1}}|\hat{x}_{V_t}) = p_{V_{t+1}}(x_{V_{t+1}}|\hat{x}^{g^{-1}}_{V_t}). \tag{4.193}$$

Now substitute $\hat{x}_{V_t} \mapsto x^g_{V_t}$ to get:

$$p_{V_{t+1}}(x^g_{V_{t+1}}|x^g_{V_t}) = p_{V_{t+1}}(x_{V_{t+1}}|x_{V_t}). \tag{4.194}$$

ad (ii): For the converse note that if Eq. (4.183) holds for an element g of a subgroup $\mathfrak{G}$ of $\Sigma_J \times \{\mathrm{i}\}$:

$$\begin{aligned}
(gp_{V_{t+1}})(x_{V_{t+1}}) &= \sum_{\bar{x}_{V_t}} p_{V_{t+1}}(x^g_{V_{t+1}}|\bar{x}_{V_t}) p_{V_t}(\bar{x}_{V_t}) && (4.195)\\
&= \sum_{\bar{x}_{V_t}} p_{V_{t+1}}(x_{V_{t+1}}|\bar{x}^{g^{-1}}_{V_t}) p_{V_t}(\bar{x}_{V_t}) && (4.196)\\
&= \sum_{\bar{x}^g_{V_t}} p_{V_{t+1}}(x_{V_{t+1}}|\bar{x}_{V_t}) p_{V_t}(\bar{x}^g_{V_t}) && (4.197)\\
&= \sum_{\bar{x}_{V_t}} p_{V_{t+1}}(x_{V_{t+1}}|\bar{x}_{V_t}) p_{V_t}(\bar{x}^g_{V_t}) && (4.198)\\
&= (P_{V_{t+1}}(gp_{V_t}))(x_{V_{t+1}}). && (4.199)
\end{aligned}$$

Where we used that we sum over all elements of $\mathcal{X}_{V_t}$ so changing the $\bar{x}_{V_t}$ to $\bar{x}^g_{V_t}$ cannot change the result.

□

Theorem 33 (Extension of symmetries of Markov matrices to the whole Markov chain)**.** *Let $\{X_i\}_{i\in V}$ be a multivariate Markov chain with $V = J \times T$ and let $\mathfrak{G}$ be a subgroup of $\Sigma_J \times \{\mathrm{i}\}$. If for all $t \in T$ a group $\mathfrak{G}$ is a subgroup of the symmetry group of the Markov matrix P_t and also a subgroup of the symmetry group of the initial distribution $p_{V_0} : \mathcal{X}_{V_0} \to [0,1]$ then for all $g \in \mathfrak{G}$*

$$gp_V = p_V \tag{4.200}$$

and $\mathfrak{G}$ is also a subgroup of the symmetry group of p_V.

Proof. First note that for all $g \in \Sigma_V$

$$gp_V = gp_{V_0,V_1,\ldots,V_T}, \tag{4.201}$$

so

$$gp_V(x_V) = gp_{V_0,V_1,\ldots,V_T}(x_{V_0}, x_{A_1}, \ldots, x_{V_T}) \tag{4.202}$$

$$= p_{V_0,V_1,\ldots,V_T}(x^g_{V_0}, x^g_{A_1}, \ldots, x^g_{V_T}) \tag{4.203}$$

$$= \prod_{t=0}^{T} p_{V_{t+1}}(x^g_{V_{t+1}}|x^g_{V_t}) p_{V_0}(x^g_{V_0}). \tag{4.204}$$

Now if $\mathfrak{G}$ is a subgroup of the symmetry group of P_{t+1} and $g \in \mathfrak{G}$ we have from Theorem 32:

$$p_{V_{t+1}}(x^g_{V_{t+1}}|x^g_{V_t}) = p_{V_{t+1}}(x_{V_{t+1}}|x_{V_t}). \tag{4.205}$$

Plug this into Eq. (4.202) to get:

$$gp_V(x_V) = \prod_{t=0}^{T} p_{V_{t+1}}(x^g_{V_{t+1}}|x^g_{V_t}) p_{V_0}(x^g_{V_0}) \tag{4.206}$$

$$= \prod_{t=0}^{T} p_{V_{t+1}}(x_{V_{t+1}}|x_{V_t}) p_{V_0}(x^g_{V_0}). \tag{4.207}$$

Since additionally $\mathfrak{G}$ is a subgroup of the symmetry group of p_{V_0} we have $gp_{V_0} = p_{V_0}$ and therefore arrive at $gp_V = p_V$. □

Remark:

- In words this theorem state that if the transition matrix commutes with the action of a group that permutes only indices within time slices (i.e. spatial indices) and the initial distribution is invariant with respect to such permutations then the joint probability distribution p_V over the entire Bayesian network is invariant with respect to such permutations.
- If the spatial permutation leaves a set $B \subset J$ fixed we can treat this as a special case of this theorem. We will do this next.

Theorem 34 (Spatial symmetries of driven multivariate Markov chains)**.** *Let $\{X_i\}_{i\in V}$ be a driven Markov chain with index set $V = J \times T$ and $J = A \cup B$ with $A \cap$*

$B = \varnothing$. Here B indicates the driving random variables and A the driven ones. Also let $\mathfrak{G}$ be a subgroup of $\Sigma_A \times \{\mathrm{i}\}$. If for all $g = (h, \mathrm{i}) \in \mathfrak{G}$, and all $x_{A,t}, x_{A,t+1}, x_{B,t}, x_{B,t+1}$ we have

$$\begin{aligned} p_{A,t+1}(x_{A,t+1}|x_{B,t+1}, x_{A,t}) p_{B,t+1}(x_{B,t+1}|x_{A,t}, x_{B,t}) \\ = p_{A,t+1}(x^g_{A,t+1}|x_{B,t+1}, x^g_{A,t}) p_{B,t+1}(x_{B,t+1}|x^g_{A,t}, x_{B,t}) \end{aligned} \tag{4.208}$$

and we are given an initial distribution $p_{A_0,B_0} : \mathcal{X}_{A_0} \times \mathcal{X}_{B_0} \to [0,1]$ with

$$g p_{A_0,B_0} = p_{A_0,B_0}, \tag{4.209}$$

then

$$g p_V = p_V. \tag{4.210}$$

Proof. First note for a driven Markov chain we have (see Definition 42)

$$p_{V_{t+1}}(x_{V_{t+1}}|x_{V_t}) = p_{A,t+1}(x_{A,t+1}|x_{B,t+1}, x_{A,t}) p_{B,t+1}(x_{B,t+1}|x_{A,t}, x_{B,t}). \tag{4.211}$$

Then we get for $g \in \Sigma_A \times \{\mathrm{i}\}$:

$$p_{V_{t+1}}(x^g_{V_{t+1}}|x^g_{V_t}) = p_{A,t+1}(x^g_{A,t+1}|x_{B,t+1}, x^g_{A,t}) p_{B,t+1}(x_{B,t+1}|x^g_{A,t}, x_{B,t}) \tag{4.212}$$

so condition Eq. (4.208) is equivalent to Eq. (4.182) for $g \in \Sigma_A \times \{\mathrm{i}\}$. This means that if

$$g p_{V_0} = g p_{A \cup B, 0} = g p_{J,0} = p_{V_0}, \tag{4.213}$$

we get $g p_V = p_V$. □

Remarks:

- If $B_t = \varnothing$ this reduces to Theorem 33.
- Here we have limited the action of the symmetry group to the driven spatial random variables. The condition just state that as long as the interaction with the driving random variables is invariant with respect to the permutations of the driven random variables, the symmetry of an initial distribution is maintained throughout the entire Bayesian network.
- A possible example system is where the drive depends on the average of all states of the nodes in A_t. Such an average is invariant under any permutation. For example let $\mathcal{X}_{B,t+1} = \{0,1\}$ and for all $j \in A_t$, $\mathcal{X}_j = \{0,1\}$ and define (for all $t \in T$:

$$p_{B,t+1}(x_{B,t+1}|x_{A,t}, x_{B,t}) = \begin{cases} 1 & \text{if } x_{B,t+1} = 1 \wedge \sum_{j \in A} x_{j,t} \leqslant |A_t|/2 \\ 1 & \text{if } x_{B,t+1} = 0 \wedge \sum_{j \in A} x_{j,t} > |A_t|/2 \\ 0 & \text{else,} \end{cases} \tag{4.214}$$

  such that $x_{B,t+1} \in \{0,1\}$ depends on whether more than half of the variables in A_t have value 1 or not. It is straightforward to check that then

$p_{B,t+1}(x_{B,t+1}|x_{A,t}, x_{B,t}) = p_{B,t+1}(x_{B,t+1}|x^g_{A,t}, x_{B,t})$ for any $g \in \Sigma_{A,t}$. The influence on the nodes in A_{t+1} also has to be symmetric, which can easily be achieved by setting all mechanisms of driven nodes (those in A) equal (just like in the case of spatial homogeneity, see Definition 39, but only among the driven nodes) e.g. let for simplicity $\mathrm{pa}(j, t+1) = (j, t)$, $\epsilon \in [0, 1]$, and set for all $j \in A$ and $x_{B,t+1}$

$$p_{j,t+1}(x_{j,t+1}|x_{B,t+1}, x_{j,t}) = \begin{cases} \epsilon & \text{if } x_{j,t+1} = x_{B,t+1} \\ 1-\epsilon & \text{if } x_{j,t+1} = x_{j,t}. \end{cases} \tag{4.215}$$

Then, if less (more) than half of the variables in A have value 1, each variable is switched to 1 (0) with probability ϵ $(1-\epsilon)$ and else stays the same. This is in effect similar to a thermostat keeping the amount of ones among the nodes in A around $|A|/2$.

Chapter 5

Agents within Markov chains

This chapter constitutes the conceptual part of this thesis. We present here an avenue for a fully formal definition of agents. Taking our cue from the literature, we extract a list of notions that, if formally defined, would suffice for a formal agent definition. We do not arrive at a full definition. For goal-directedness, which would complete such a definition we make no proposal. For the other notions we propose definitions and motivate these in each case. Finally, we connect the resulting proto-agents (lacking goal-directedness) to the existing formal model of agent-environment system called the perception-action loop.

In more detail the chapter contains the following.

- In Section 5.1 we give a working definition of agents in accordance with the literature, this stipulates agents as entities contained within systems, the entities must exhibit actions, perception, and goal-directedness.
- In Section 5.2 we restrict ourselves to a class of target systems that we want to define agents for, i.e. (driven) multivariate Markov chains.
- In Section 5.3 we present an arguments for three phenomena that should be possible (and not be precluded) under a general agent definition: compositionality, degree of freedom traversal, and counterfactual variation. We also propose spatiotemporal patterns (STPs) as structures that can exhibit these three features within multivariate Markov chains.
- In Section 5.3.5 we highlight the problem of selecting entities among all STPs i.e. which we call the problem of identity. We also propose completely locally integrated STPs as one possible solution.
- In Sections 5.4 and 5.5 we present definitions of actions and perceptions that are suitable for any notion of entities that is based on STPs.
- In Section 5.6 we show that our proposed notions of action and perception can be seen as generalisations of existing notions in the perception-action loop literature.

The original contributions contained in this chapter are:

- An argument (via compositionality, degree of freedom traversal, and counterfactual variation) for a STP-based definition of entities.
- The abstraction of entity-sets which enables the formal connection to perception-action loop.
- A tentative[1] formal definition of entities as completely locally integrated STPs.
- A tentative formal definition of action for arbitrary entity-sets.
- A classification of actions into value actions and extent actions.
- A tentative formal definition of perception for arbitrary entity-sets.
- An exposition of the role of non-interpenetration of entity-sets in perception. Namely, it makes perception naturally unique.
- The formal exposition of the connection of the action definition to non-heteronomy of Bertschinger et al. (2008) in the perception-action loop.
- The formal exposition of the way the perception definition specialises to the perception-action loop.
- A construction of a conditional probability distribution (the branch-morph, including branching partition) over the futures of entities which allows the definition of perception.
- Proof that the condition on co-perception environments is not stronger than the assumptions about environment states inherent in the perception-action loop.

The formal definitions of entities, actions, and perceptions are only tentative. We will establish in this thesis that they are compatible, i.e. the defined actions and perceptions apply to the defined entities. We also establish in Sections 6.5 and 6.6 that they fulfil certain expectations and the requirements we propose in this thesis. They remain tentative for three reasons (in increasing generality):

1. There are still some open questions regarding their interpretation in certain cases. See Sections 6.5 and 6.6 and Chapter 7.
2. There are further phenomena that should not be precluded by agent definitions that we have not investigated yet such as death, birth, growth, and replication.
3. A final formal definition of agents should also be empirically grounded. For this it should correctly predict the conditions for the occurrence of agents in some system. For this it is in turn necessary that the occurring agents are more or less universally acknowledged examples of agents.

[1] For some context on what we mean by "tentative" see Chapter 5.

The best examples of such agents are humans but some researchers are willing to attribute agency to much simpler living organisms like bacteria. In that case one can imagine that a formal definition of agents could in principle get empirical justification: Say we can formally model beakers filled with suitable chemicals and their dynamics. A formal agent definition would be empirically justified if it can be used to reliably predict the conditions (external influence and initial conditions) that lead to the occurrence of agents e.g. the emergence of a bacterium within the beaker. This scenario can (in principle only) also be extended to the case where only humans are agents. In practice this seems further magnitudes more unlikely than the scenario involving the bacterium.

Another and possibly simpler path to recognition of an agent definition would be if it was able to predict which formal systems that can be simulated contain agents and where such simulations turn out to be "convincing enough" to the research community. Currently neither of these scenarios has been realised with our formal agent definition. Neither scenario is also likely to happen in the near future. A classification of formal systems according to their capacity to contain agents will require not only further mathematical tools but also the right guess with respect to the agent definition. This thesis presents one such guess and develops some associated new mathematical tools. Whether either proposal will be part of a final and acknowledged formal definition of agents is work for the future.

5.1 A working definition of agents

Conceptually, agent definitions are not particularly controversial in the literature. As Barandiaran et al. (2009) have argued a rough consensus is that an agent is "at least, *a system doing something by itself according to certain goals or norms within a specific environment*" (Barandiaran et al., 2009, p.2, italics from the original). In their subsequent discussion they highlight the necessity for a "distinguished entity" to exist in order to take the role of the "system doing something". For the moment this almost suffices for our purposes. It is not the main goal of this work to question the concept of agents but instead to contribute to its formalisation within a preselected class of formal/artificial systems.

To get a concise working definition we insert the "distinguished entity" into the consensus definition above and reformulate it slightly:

Definition 64 (Agent working definition)**.** *An agent is a distinguished entity contained within a strictly larger system exhibiting perception, action, and goal-directedness.*

Instead of referring to a "specific environment" we only require a larger system that encompasses/contains the agent as well as something more. The "something more" can certainly take the form of an environment and, conversely, if there is an environment the "larger system" can always be defined

as the agent together with its environment. If anything our definition is therefore more general than that of Barandiaran et al. though this is not the main purpose of this formulation. The main purpose is to highlight the existence of the larger system which *contains* the agent entirely. It is important to remember that this existence is not a new requirement but is present already in the prevalent concept of agents via the reference to an "environment". Also note that when we require that the distinguished entity is contained within a larger system, we mean the entity cannot be *everything that exists* for any amount of time. By "larger" we therefore mean the system is larger during and throughout the agent's existence/presence and not "larger" only in the temporal domain.

It is worth highlighting two more things about how we interpret our working definition. The first is that an agent is a "distinguished entity" in so far that it exhibits actions, perception, and goal-directedness. These three properties *distinguish* an entity that is an agent among all entities. This leads directly to the second point: the set of agents is a (set theoretical) *subset* of the set of entities contained in the larger system. From this it follows that the set of entities has to be defined in such a way that it encompasses all agents. We will come back to this in Section 5.3.

That we choose to use "actions" instead of "doing something by itself" and "goal-directedness" instead of "according to certain goals or norms" should not be over-interpreted. We consider these expressions as interchangeable. We also included perception in the definition because it is (sometimes via "interaction") a very common requirement in the literature (e.g. Maes, 1993; Beer, 1995; Smithers, 1995; Franklin and Graesser, 1997; Christensen and Hooker, 2000). Furthermore, Barandiaran et al. (2009) in their more detailed discussion of the requirements for agency refer to "interactional asymmetry" instead of "doing something". The use of interaction suggests that they also agree with a requirement of perceptions.

We also want to draw attention to the fact that Definition 64 is still a very weak definition of agents. Especially when discussing biological agents i.e. living organisms, further requirements are common. One such requirement concerns the relation between the goals and the agent. In the definition above this relation is arbitrary, any goals are valid for any agent. Stronger definitions require the goals to be somewhat intrinsic to the agent. This can mean that goals must be in the agent's own interest (e.g. Franklin and Graesser, 1997; Kauffman, 2000) for example ensure its survival/existence (cf. constitutive autonomy e.g. Froese and Ziemke, 2009).

Primary examples of agents are living organisms, the higher organised they are the less controversial the claim that they are indeed agents. In the end it cannot be denied, for example, that humans are agents. Many authors agree that bacteria already qualify as (sometimes called minimal) agents (Christensen and Hooker, 2000; Kauffman and Clayton, 2006; Froese and Ziemke, 2009; Barandiaran et al., 2009). Due to the supposed lack of representational capabilities others do disagree (see Schlosser, 2015, for references). We take the point of view in line with our definition of agents above that bacteria and all living organisms are agents. For the purpose of this thesis it is not essential

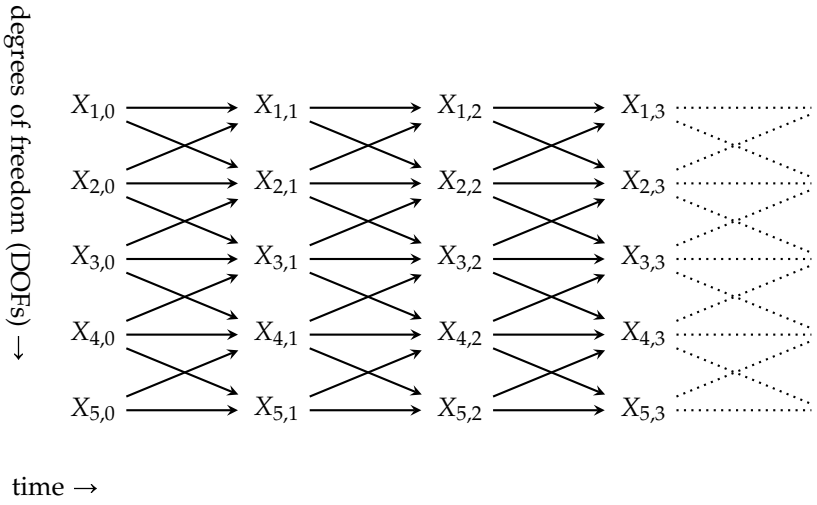


Figure 5.1: First time steps of the Bayesian network representing a multivariate Markov chain $\{X_t\}_{t\in T}$. The shown edges are just an example, any two nodes within the same or subsequent columns can be connected.

to make a final commitment on these matters. It would suffice for our arguments if we would use a more restrictive definition, e.g. that only humans are agents. Whenever we speak of living organisms in the following the inclined reader might then just replace this by "humans" and the same or similar arguments still hold. Apart from living organisms, other examples of agents are robots, and more controversially, societies, companies, and nation states.

5.2 Multivariate Markov chains as a class of systems containing agents

In order to transform the working Definition 64 into a formal definition we have to formally define every one of the terms mentioned there. The most fundamental term in Definition 64 is the "larger system". The distinguished entities must be "parts of" the larger system so to define those we need to define the larger system first. Perception, action, and goal-directedness can then be defined once we have well defined distinguished entities.

As the class of larger systems we choose finite multivariate Markov chains (see Definitions 36 and 42 as well as Figs. 5.1 and 5.2).

One immediate consequence of choosing a well defined class of systems is that it forces us to construct all other notions from those well defined for this class of systems. This has the advantage that it greatly restricts the concepts to consider for the definitions. The disadvantage is that if the choice is a bad choice we are destined to fail. A bad choice here would mean that a useful notion of agents is impossible within our choice of larger systems. In this section we therefore explain what motivates our choice of finite Markov

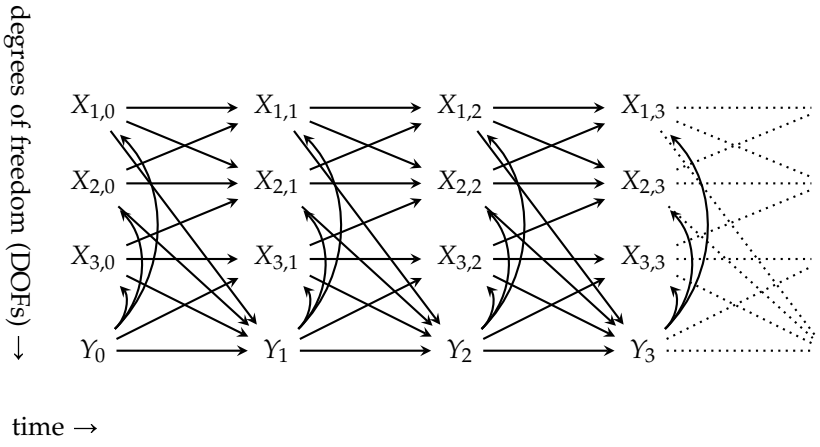


Figure 5.2: First time steps of the Bayesian network representing a multivariate process $\{X_i\}_{i\in V}$ driven by a process $\{Y_t\}_{t\in T}$. Note that the process $\{Y_t\}_{t\in T}$ can also be multivariate, but this would further clutter the graph. Also note that not all edges depicted here must be present. Here, each random variable in each time-slice of the driven process is influenced by the driving process and influences it.

chains as the "larger systems".

Since we have required that agents are parts of larger systems in Definition 64 this system must in one way or another *contain* the agent candidates. As a choice for the class of systems that represent the larger systems we should then use systems for which it is plausible that they can contain agents. At the same time want to start with the *simplest* class of systems that shows at least some promise or is not easily dismissed.

This desire for simplicity is due to two factors. First, in artificial life we are mainly interested in the principles that allow the occurrence of agents/life within a system and not in the precise description of actual agents/living organisms. Second, choosing simpler systems greatly reduces the technical burden so that the concepts play a more prominent role.

The main factor in choosing a finite system is that implementations of the systems under consideration in computer simulations are of great interest in artificial life. Such simulations are restricted to discrete and finite systems. Continuous systems can be approximated, but the approximations are in the end finite again so such approximations are included in the class of finite systems.

As living organisms are our prime example of agents it would be straightforward to choose systems which resemble or model systems that contain living organisms. The safest bet is then to use models of the universe as a whole. By definition the universe contains living organisms entirely.

Another reasonably safe bet is the entire geosphere, by which we mean the planet earth together with its atmosphere and the exchange of radiation with the sun and the rest of the universe. Similarly there are smaller subsystems of the geosphere that can contain living organisms like ponds, tidal pools, and

other ecosystems.

Realistic models of the universe or the geosphere are continuous, use quantum mechanics (e.g. as in Saitta and Saija, 2014), relativistic mechanics, or even more involved theories.

However, for the sake of simplicity and finiteness we abstract away from the more realistic continuous, relativistic, or quantum dynamical systems to finite multivariate Markov chains Definitions 36 and 42.

This class of systems contains synchronous finite cellular automata like the game of life (Conway, 1970). These automata can be seen as discretised versions of field theories (Shalizi and Moore, 2003) and have successfully been used to model physical systems Chopard (2009).

Multivariate Markov chains can also be used to approximate particle-based systems if we use the random variables to represent the positions and momenta of the particles. An interesting recent system with life-like behaviour which falls into this class is Schmickl et al. (2016).

Since we also include driven multivariate Markov chains our considerations also extend to reaction-diffusion systems. Such systems are also frequently used to model biological phenomena (Turing, 1952), as well as individualised and metabolising structures (Virgo, 2011; Froese et al., 2014; Bartlett and Bullock, 2015, 2016).

Last but not least, the driven multivariate Markov chains can be used to approximate/simulate systems obeying the (multivariate) Langevin equation (as a discretised version of the associated Fokker-Planck equation). Such systems underlie recent investigations into the physics of cell replication (England, 2013) and adaptation (Perunov et al., 2014) as well as a theory of life (Friston, 2013). They are also used in Still et al. (2012) to study advantages of prediction, perception and action for thermodynamic efficiency. Similarly Sagawa (2012) studies the increased work extraction due to feedback control, which can be seen as perception and action as well. Kondepudi et al. (2015) studies a form of goal-directedness in a driven system.

Therefore we believe that multivariate (driven) Markov chains are a reasonable choice for a first class of systems to develop an agent definition for. Our hope is that these system are powerful enough to contain agents. However, if we find that they are not then we may at least find out why they are not. At the current state of this research this question is still undecided.

5.3 Entities in multivariate Markov chains that can be agents

As mentioned in Section 5.1 all agents are entities. After choosing multivariate Markov chains as the class of containing systems we have to define entities in this class of systems. For this purpose we propose to use subsets of STPs in general and completely integrated STPs (Definition 53) in particular. Employing STPs to represent entities is already implicit in (Beer, 2014b,a). The notion of completely integrated STPs and the proposal of using them as entities in an agent definition are two of the main original contributions of this thesis. The

section is loosely based on our own publication Biehl et al. (2016).

Formally, the first goal of this section is to establish that the set of entities $\mathfrak{E}(\{X_i\}_{i\in V})$ for any given (driven) multivariate Markov chain[2] $\{X_i\}_{i\in V}$ should be a subset of the STPs of $\{X_i\}_{i\in V}$ i.e.

$$\mathfrak{E}(\{X_i\}_{i\in V}) \subseteq \bigcup_{O\subseteq V} \mathcal{X}_O. \tag{5.1}$$

Using subsets of STPs is in contrast (i.e. *not* equivalent) to using subsets of random variables i.e.

$$\mathfrak{E}(\{X_i\}_{i\in V}) \subseteq 2^V \tag{5.2}$$

as entities. The latter are often implicitly used in the literature. The arguments for Eq. (5.1) are to a large degree independent of arguments that concern the exact determination of *which* subset $\mathfrak{E}(\{X_i\}_{i\in V})$ should correspond to. Consequently, there may be different notions of agents based on different choices of the exact subset. In order to accommodate this we introduce the notion of *entity sets* which later allow us to define actions and perception independent of the exact choice of $\mathfrak{E}(\{X_i\}_{i\in V})$. The exact choice is the subject of the second part of this section and the problem of identity. There we will motivate the choice of completely integrated STPs

$$\mathfrak{E}(\{X_i\}_{i\in V}) = \{x_O \in \bigcup_{O\subseteq V} \mathcal{X}_O : \iota(x_O) > 0\}. \tag{5.3}$$

Note that this section does not contain a rigorous derivation of the necessity to choose the entities as we propose. We merely present heuristic arguments which speak for this choice. The main tool in this endeavour is the following argument.

As already mentioned in Section 5.1 the set of entities for a given larger system has to (at least) encompass all agents within the system. In other words the definition of entities must not exclude structures which might be agents. Now say that there are phenomena or properties that are known to be exhibited by some (possibly not all) agents. Say that furthermore there is an entity definition which implies that these phenomena or properties are *impossible* for entities. Then we must reject this entity definition on the grounds that it cannot encompass all agents since it *precludes* these phenomena or properties. In the following this argument will be employed multiple times and referred to as the *non-preclusion argument*. Note that we cannot require all phenomena that are exhibited by some agents to be exhibited by all entities of an entity definition. This would lead to a small and possibly empty subset of agents. We can however require that all phenomena that are exhibited by some agents are not-precluded by the entity definition. In this way every phenomenon that is exhibited by some agent can be turned into a condition on entity definitions.

[2]We will not explicitly mention "driven" in the following. We will also refer to the Markov chain $\{X_i\}_{i\in V}$ without explicitly mentioning driving or driven random variables. The process $\{X_i\}_{i\in V}$ should be seen as the process of interest that may or may not be driven by some other process whose dynamics are ignored. For the purpose of this chapter whether $\{X_i\}_{i\in V}$ is driven or not makes no difference.

We illustrate our arguments for choosing entities within the class of multivariate Markov chains using the popular example of a glider in the game of life cellular automaton. The glider is not necessarily a life-like structure, but it already exhibits the three phenomena that we will further discuss in this thesis:

1. compositionality,
2. degree of freedom traversal,
3. counterfactual variation.

These phenomena are also exhibited by more life-like structures in less well known examples of (driven) multivariate Markov chains. We refer the reader to the motile and interacting reaction-diffusion spots in Virgo (2011); Froese et al. (2014), different reaction-diffusion spots in Bartlett and Bullock (2015, 2016), and the particle-based cell-like structures in Schmickl et al. (2016). In the following when we refer to "other life-like structures" we refer to these examples.

Note that there are further phenomena of living organisms and life-like systems that should not be precluded by an entity definition. Examples of such phenomena are birth, death, growth, and replication. The investigation of these is beyond the scope of this thesis.

In Sections 5.3.1 to 5.3.3 we discuss each of the three phenomena above separately. For each we will also note that they seem plausible for real living organisms. We then invoke in each case the non-preclusion argument and require that a definition of entities in multivariate Markov chains should allow structures that exhibit this phenomenon. In the course of these arguments we settle for STPs as the superset of entities and state this explicitly in Section 5.3.4. There we will also define compositionality, degree of freedom traversal, and counterfactual variation formally. This leaves open the problem of selecting entities among all STPs which is the problem of identity discussed in Section 5.3.5.

5.3.1 Compositionality of entities

The compositionality of entities refers to the possibility that life-like structures are composite of multiple parts. This can be separated into two kinds of compositionality, spatial compositionality and temporal compositionality. We first discuss spatial compositionality.

In the example of the glider we observe that a glider is not just a single cell or the state of a single cell (e.g. black or white). In order for a glider to occur multiple cells that are in a particular arrangement have to have particular states at some time-step. The glider is therefore a (spatially) composite structure. The same is true for other life-like structures.

In reaction-diffusion systems (e.g. Froese et al., 2014; Bartlett and Bullock, 2015) the individualised spots occupy a contiguous bounded region in a two dimensional plane. Each position in the reaction-diffusion system is a random

variable that indicates the concentrations of the involved chemicals at this position. A single position or a single set of concentrations at a position does not constitute a spot or life-like structure these are composite of all the concentrations in an area. So reaction-diffusion spots are also spatially composite structures.

In the particle-based system of Schmickl et al. (2016) the life-like structures are spores or cells. These are composite of particles. In this case each particle j has three degrees of freedom, two positions $x_{j,t}, y_{j,t}$ and heading $\phi_{j,t}$. Each degree of freedom can represented by a random variable. The union of these degrees of freedom over all particles and all times form the random variables of the multivariate Markov chain. A spore or cell occurs if these random variables stand in particular relation to each other. Without going into further details, a necessary condition is that the positions of multiple particles must be (in some sense) close to each other. A single particle is not a cell or a spore so again we find that life-like structures are spatially composite structures.

Finally, living organisms (presumably composite of molecules) are also generally seen as composite structures.

So by the non-preclusion argument the entities in multivariate Markov chains should include spatially composite structures.

The glider is also a temporally composite structure. It is an essential feature of a glider that it "moves" which means that in general it can exist at multiple time-steps and we refer to it as the same glider. This means the glider can persist or can be composite out of parts at different time-steps.

Similarly, the life-like structures in other systems are persistent structures. Furthermore living organisms persist and are generally seen to have histories which is another indication that they are composite out of parts at different times. In Section 5.3.2 we will discuss temporal compositionality with particular attention to the possibility that the spatial parts that the glider or many life-like structures are composite of change over time. We therefore keep the discussion of temporal compositionality short.

Note that both choices of entities $\mathfrak{E}(\{X_i\}_{i\in V}) \subseteq 2^V$ and $\mathfrak{E}(\{X_i\}_{i\in V}) \subseteq \bigcup_{O\subseteq V} \mathcal{X}_O$ can represent composite structures in the form of sets of random variables or sets of values of random variables respectively.

5.3.2 Degree of freedom traversal of entities

Degree of freedom traversal refers to the possibility that life-like structures maintain a form of identity while exchanging the spatial parts they are made of.

The glider in the game of life "moves" in one of four possible directions. As it moves the cells that it occupies change at every time step. Nonetheless we speak of the *same* glider even when none of the cells it occupied in one configuration are still occupied several time steps later. Intuitively then the glider maintains its identity along its path. We call the maintenance of identity under exchange of the spatial occupied cells or spatial occupied random variables *degree of freedom traversal*. Note that a block in the game of life does not exhibit such degree of freedom traversal.

Other life-like structures in multivariate Markov chains also exhibit degree of freedom traversal. In reaction-diffusion systems (e.g. Froese et al., 2014; Bartlett and Bullock, 2015) the individualised spots travel through the two dimensional plane and thereby occupy changing spatial regions over time. As mentioned before, each position in a reaction diffusion system indicates a random variable representing the concentrations of the involved chemicals at this position. So reaction-diffusion spots also exhibit degree of freedom traversal.

In the particle-based system of Schmickl et al. (2016) particles can be seen to jump into and out of the spores and cells which intuitively maintain their identity throughout. As mentioned before each particle corresponds to three degrees of freedom/random variables. The "spatial" random variables that are occupied by a spore or cell are those of the particles that it is formed by. This means that when a particle jumps in or out of the spore or cell the random variables occupied by this structure change. Therefore these structures also exhibit degree of freedom traversal.

Furthermore, we also see real living organisms as maintaining their identity while exchanging the parts they are made of. The molecules that a cell is made of change during its lifetime. So if a cell (or a larger living organism) is seen as a particular configuration of molecules then these also exchange the parts they are made of.

Together these observations suggest that degree of freedom traversal is exhibited by life-like structures in Markov chains. According to the non-preclusion argument a definition of entities should therefore allow the possibility of entities that traverse degrees of freedom.

Both candidates for entities mentioned in Section 5.3.1 can represent degree of freedom traversal. The subsets of random variables $\mathfrak{E}(\{X_i\}_{i\in V}) \subseteq 2^V$ also contain sets of random variables that differ from time-step to time-step. This is possible because we use the "time-unrolled" Bayesian network formulation of multivariate Markov chains where each time step has its own set of random variables $\{X_i\}_{i\in V_t}$ for time-slice V_t. A subset of V can then combine any subsets of the time-slices at different times.

The subsets of STPs $\mathfrak{E}(\{X_i\}_{i\in V}) \subseteq \bigcup_{O\subseteq V} \mathcal{X}_O$ clearly also contain STPs that fix different random variables at different times since they can fix the random variables in any subset of V.

We note here again that the choice of entity sets $\mathfrak{E}(\{X_i\}_{i\in V}) \subseteq 2^V$ was used in Krakauer et al. (2014) to construct a notion of what is called individuals in this work. This work also deals with degree or freedom traversal of these individuals. However, the phenomenon of life-like structures discussed in Section 5.3.3 suggests that this construction precludes certain structures that we would like to include. This is also the case for the perception-action loop (Section 3.3.6) where the agent is represented by the set of random variables $\{M_t\}_{t\in T}$.

5.3.3 Counterfactual variation of entities

Counterfactual variation of structures refers to the possibility that the life-like structures within a multivariate Markov chain differ from one *trajectory* to another. The difference or variation can take two forms:

1. variation in value,
2. variation in extent.

We could add "variation in existence" but this is a special case of variation in extent. Before we give a formal definition of counterfactual variation we can already state a few observations about gliders. First, a glider can occur in one trajectory of the cellular automaton and not occur in a second one (e.g. if all cells are white in the second one). This would be a variation in existence. It can be seen as a variation in extent since its extent in the second trajectory is zero.

Second, a glider can occur in one place and move in one direction in the course of one trajectory and occur in another place and move in another direction in the course of a second trajectory. This is also a variation in extent if the cells that the two gliders occupy in the two trajectories along their path do not completely coincide.

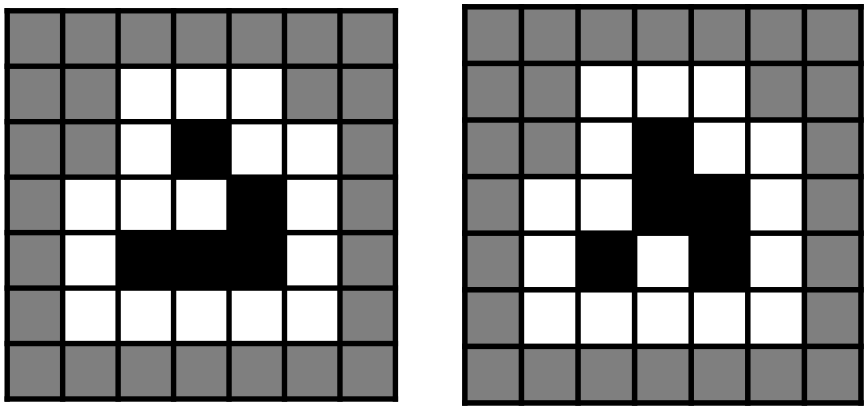
Third, a variation in value occurs if in two trajectories the structures/gliders occupy exactly the same cells but the cells they occupy have differing values. For gliders this can happen if we look at trajectories that are only one time-step long Fig. 5.3.

Note that if two glider in two trajectories have the same values and extent we consider it them the same glider.

Note that it is a non-trivial question which cells/random variables the glider actually occupies. A detailed discussion can be found in Beer (2014b). The example of counterfactual variation in value in Fig. 5.3 is based on the characterisation of a glider as a STP by Beer.

We have seen that gliders are structures that exhibit counterfactual variation of the two kinds. For real living organisms it is impossible to say with certainty whether they exhibit counterfactual variation. We have no access to counterfactual trajectories of the universe. However, due to the symmetries of the laws of physics it is quite plausible that living organisms also exhibit counterfactual variation. Assume we have two identical aquarium containing a different fish each. The laws of physics suggest that the fish in one aquarium could also be in the same place in the other aquarium and similarly for the other fish. This would be counterfactual variation in extent of the fish within either aquarium.

Together with evidence from other life-like structures we therefore consider it justified to invoke the non-preclusion argument and require that counterfactual variation is possible for entities within multivariate Markov chains. Therefore it should be possible for entities to occur within single trajectories and not occur in others. This suggests STPs as candidates for entities i.e. $\mathfrak{E}(\{X_i\}_{i\in V}) \subseteq \bigcup_{A\subseteq V} \mathcal{X}_A$.



(a) (b)

Figure 5.3: Counterfactual variation in value exhibited by the spatial patterns of two gliders in two different trajectories of a game of life cellular automaton. Both grids depict a single time-step of the two dimensional cellular automaton and therefore a special case of STPs without temporal extension. The cells that are not occupied by the patterns are grey, occupied cells are white or black according to the glider configuration. The extra layer of white cells around the black cells of the gliders is in accordance with the extent of gliders as derived by Beer (2014b). The two particular configurations shown here are also presented there. We see that the cells that are occupied coincide in both cases. The only difference between the two glider configurations are the values of two cells. The one right in the centre of the grid and the cell just below it have switched their values. This shows that there can be a glider in one trajectory and a different glider in another trajectory with both having identical extent. This is a counterfactual variation in value.

Conversely, this throws out the possibility that entities are *by definition* structures that do not vary with the trajectories. In particular this excludes the possibility to define entities as random variables, sets of random variables, or stochastic processes i.e. $\mathfrak{E}(\{X_i\}_{i\in V}) \notin 2^V$ since these do not vary with the trajectories. A subset of random variables X_A where $A \subseteq V$ is not dependent on the trajectories, it only takes different *values* in different trajectories. The different values are just STPs however. Note that this does not mean that it is not justified in *particular* cases to represent entities by sets of random variables. In particular, if it happens that every STP $x_A \in \mathcal{X}_A$ is an entity then it could make sense to speak of these STPs as different realisations of one thing / an individual / an "entity" [3]. For example, this is the assumption for the agent process $\{M_t\}_{t\in T}$ in the perception-action loop (see Section 3.3.6) and will be discussed further in Sections 5.5 and 5.6.

We then settle for STPs as the candidates for entities i.e. entities are seen as special cases of STPs. For a given multivariate Markov chain entities are therefore a subset of the STPs. We fix this assumption formally in the next section where we also give a formal definition of compositionality, degree of freedom traversal, and counterfactual variation.

5.3.4 Definition of entity sets via STPs

In accordance with Section 5.3.3 we define the set of entities in a multivariate Markov chains to be a subset of the STPs. We call this subset the entity set.

Definition 65 (Entity set)**.** *Given a multivariate Markov chain $\{X_i\}_{i\in V}$ with index set $V = J \times T$ an entity set $\mathfrak{E}(\{X_i\}_{i\in V})$ is a subset of the set of all STPs $\bigcup_{A\subseteq V} \mathcal{X}_A$ i.e.*

$$\mathfrak{E}(\{X_i\}_{i\in V}) \subseteq \bigcup_{A\subseteq V} \mathcal{X}_A. \tag{5.4}$$

- We expect that it is useful to require certain algebraic properties from entity sets. However, this is beyond the scope of this thesis.

Using entity sets we can also define compositionality, degree of freedom traversal, and counterfactual variation formally. To connect this with the example of a glider recall that a trajectory $x_V \in \mathcal{X}_V$ of a multivariate Markov chain is a STP that extends throughout the entire Bayesian network. In other words a trajectory occupies all random variables in the Markov chain. The multivariate Markov chain describing a cellular automaton assigns each cell $j \in J$ (where J is a two dimensional grid of cells) at each time step $t \in T$ a random variable $X_{j,t}$. The random variables in the multivariate Markov chain are then indexed by $V = J \times T$ and a trajectory is a STP $x_V = x_{J,T}$. Structures like the glider (and similar life-like structures) occupy/fix subsets of the random variables in the multivariate Markov chain. They can therefore be described by STPs.

[3] The quotations only indicate that we will reserve the term *entity* for STP based entities in the rest of this thesis.

We can then define spatial and temporal compositionality, first for STPs, then for entity sets.

Definition 66 (Composite STPs)**.** *Given a multivariate Markov chain $\{X_i\}_{i\in V}$ we say a STP x_A is*

(i) spatially composite or extended if it has a time-slice x_{A_t} occupying more than one random variable i.e. if there exists t with $|A_t| > 1$,
(ii) temporally composite or extended if it has more than one non-empty time-slice i.e. if $|\{t \in T : A_t \neq \varnothing\}| > 1$.

If x_A is spatially and temporally extended we say it is spatiotemporally composite or extended.

Definition 67 (Compositionality)**.** *Given a multivariate Markov chain $\{X_i\}_{i\in V}$ the entity set $\mathfrak{E}(\{X_i\}_{i\in V})$ satisfies compositionality if it contains a composite STP.*

Similarly for degree of freedom traversal.

Definition 68 (Degree of freedom traversing STPs)**.** *Given a multivariate Markov chain $\{X_i\}_{i\in V}$ with index set $V = J \times T$ we say a STP x_A traverses degrees of freedom if there are two time-slices that occupy random variables with different spatial indices i.e. if there exists $t, s \in T$ with $t \neq s$ such that*

$$\{j \in J : (j,t) \in A_t\} \neq \{j \in J : (j,s) \in A_s\}. \tag{5.5}$$

Definition 69 (Degree of freedom traversing entity set)**.** *Given a multivariate Markov chain $\{X_i\}_{i\in V}$ the entity set $\mathfrak{E}(\{X_i\}_{i\in V})$ is degree of freedom traversing if it contains a degree of freedom traversing STP.*

Remark:

- The indices of a multivariate Markov chain may be renamed such that the property of degree of freedom traversal vanishes from the entity set. We accept this caveat here. If needed the notion of degree of freedom traversal can be strengthened by requiring that no such index renaming removes the property. Note that if there are two entities, one that doesn't traverse degrees of freedom and one that does and both occupy the same degree of freedom at some time t then they make it impossible to rename the indices at all times where the two entities differ. This is the case for the two gliders of Fig. 5.3. Also, since the renaming of indices is rarely practically done or considered in case of cellular automata (it leads to complex update rules) and other systems exhibiting life-like phenomena our simple notion of degree of freedom traversal is sufficient for the purpose of this thesis.

Finally, we define counterfactual variation by first defining variation (or difference) in value and extent:

Definition 70 (Variation of STPs)**.** *Given a multivariate Markov chain $\{X_i\}_{i\in V}$. Two STPs $x_A, \bar{x}_B$ differ or vary*

(i) in value if $A = B$ and there exists $i \in A$ with $x_i \neq \bar{x}_i$,

(ii) in extent if $A \neq B$.
(iii) in value and extent if $A \neq B$ *and there exists* $i \in A \cap B$ *with* $x_i \neq \bar{x}_i$.

We just say $x_A, \bar{x}_B$ *differ or vary if any of the above are true. Else we say they are identical or equal.*

Remark:

- We will encounter the difference in value and extent again when we define actions for agents based on STPs in Section 5.4. There we can distinguish between actions in value and actions in extent.

We then have counterfactual variation if the set of all entities in one trajectory differs from the set of entities in another.

Definition 71 (Counterfactually varying entity set)**.** *Given a multivariate Markov chain* $\{X_i\}_{i\in V}$ *the entity set* $\mathfrak{E}(\{X_i\}_{i\in V})$ *exhibits*

(i) counterfactual variation if there are two trajectories $x_V, \bar{x}_V \in \mathcal{X}_V$ *with* $x_V \neq \bar{x}_V$ *such that the set of entities that occur in each are not equal i.e.*

$$\{\hat{x}_A \in \mathfrak{E}(\{X_i\}_{i\in V}) : \hat{x}_A = x_A\} = \{\hat{x}_A \in \mathfrak{E}(\{X_i\}_{i\in V}) : \hat{x}_A = \bar{x}_A\} \tag{5.6}$$

(ii) counterfactual variation in value only if for any two trajectories $x_V, \bar{x}_V \in \mathcal{X}_V$ *set of entities that occur in each only differ in value which means that they all occupy the same sets of random variables i.e.*

$$\{A \subseteq V : \exists \hat{x}_A \in \mathfrak{E}(\{X_i\}_{i\in V}), \hat{x}_A = x_A\} = \{A \subseteq V : \exists \hat{x}_A \in \mathfrak{E}(\{X_i\}_{i\in V}), \hat{x}_A = \bar{x}_A\} \tag{5.7}$$

(iii) counterfactual variation in extent if it exhibits counterfactual variation but not counterfactual variation in value only.

In Section 6.5 we will see that the entity set we propose in Section 5.3.5.2 exhibits all three of these phenomena. Next we turn our attention towards choosing the right entity set from among all STPs.

5.3.5 The problem of identity

5.3.5.1 General considerations

Roughly, the problem of identity is the problem of determining which structures within a system form a (possibly) composite entity and which structures don't. We have already mentioned the maintenance of identity that we attribute to gliders and other life-like systems in Section 5.3.3. Since we have now decided on entities as subsets of STPs we can now look at identity more closely.

We can also state the problem of identity formally.

Definition 72. *Given a multivariate Markov chain* $\{X_i\}_{i\in V}$ *the problem of identity is the problem of deciding on a particular entity set* $\mathfrak{E}(\{X_i\}_{i\in V})$.

This immediately suggests the trivial solutions of choosing all STPs as entities:

$$\mathfrak{E}(\{X_i\}_{i\in V}) = \bigcup_{A\subseteq V} \mathcal{X}_A. \tag{5.8}$$

This trivial solution is akin to what is called "unrestricted mereological composition" in philosophy Gallois (2015). Every combination of spatial, temporal, and STPs is an entity (or object in the case of unrestricted mereological composition). This includes all trajectories whether they are possible under the mechanisms of the multivariate Markov chain or not.

However, our intuition is that some STPs are more entity-like than others. We think that a glider or a block in the game of life is more of an entity than, for example, half of a glider together with half a block far away in the grid. We also think that a glider pattern at time t and together with the subsequent glider pattern at time $t+1$ are more of an entity than the glider at time t and the far away block at time $t+1$ taken together as one STP. Similarly, in the real world we also see differences between the degrees to which certain subsets of the world are entities. An animal's leg together with the rest of its body forms more of an entity than its leg together with part of a nearby (or indeed far away) tree trunk. Living organisms account for a large number of examples but also some non-biological structures at least seem like entities e.g. hurricanes, tornadoes, etc., (soap) bubbles, and maybe human created artefacts.

The question is then what makes some composite structures more "entity-like" than others. An answer would be that there is a special relation that holds between the parts of "entity-like" structures that does not hold (or holds to a lesser degree) for other structures. In the case of STPs in multivariate Markov chains we should then look for relations between the parts of the STPs. Here, the different parts of STPs are related (or unrelated) due to the dynamics of the system. These dynamics are defined via the mechanisms of the multivariate Markov chain and generate the probability distribution over the entire chain. In this sense the probability distribution over the chain contains all information about relations between STPs (and their parts since the parts are again STPs). This suggests using the probabilities to formulate a quantitative condition, relation, or measure that given a STPs tells us in how far it constitutes and entity.

5.3.5.2 Completely locally integrated STPs as entities

There are certainly multiple candidates for such a measure. It is beyond the scope of this thesis to provide a systematic comparison between multiple such candidates. We only propose complete local integration as on instance of such a measure. The formal structure of complete local integration has been investigated in Chapter 4 and we provide examples of this structure in Chapter 6. In future work we will investigate other identity measures and also plan to try more axiomatic approaches. A starting point for an axiomatic approach would be the questions raised in the philosophical discussion on location and serology (Gilmore, 2014). One such question is whether entities should be

allowed to overlap or interpenetrate[4]. Choosing an answer could be turned into an axiom which might restrict the possible identity measures. However here we only present some intuitions behind the first candidate for an identity measure: complete local integration.

For this let us treat the problem of identity as a combination of

1. spatial identity and
2. temporal identity

for the moment. In the end we will propose a solution which makes no distinction between these two aspects. We note here that conceiving of entities (or objects) as composite of spatial and temporal parts as we do in this thesis is referred to as four-dimensionalism or perdurantism in philosophical discussions (see e.g. Hawley, 2015). The opposing view holds that entities are spatial and endure or persist over time. This view is called endurantism. Here we will not go into the details of this discussion.

The main intuition behind complete local integration is that every part of an entity should make every other part more probable.

This seems to hold for example for the spatial identity of living organisms. Isolated parts of living organisms are *rare* compared to whole living organisms. For example it is rare to see only an arm without the rest of a human body attached compared to seeing an arm with the rest of a human body attached. The body seems to make the existence of the arm more probable and vice versa. This seems to hold for all living organisms but also for some non-living structures. The best example of a non-living structure we know of for which this is obvious are soap bubbles[5]. Half soap bubbles (or thirds, quarters,...) only ever exist for split seconds whereas entire soap bubbles can persist for up to minutes. Any part of a soap bubble seems to make the existence of the rest more probable. Similarly, parts of hurricanes or tornadoes are rare. So what about spatial parts of structures that are not so entity-like? As a crude approximation we can think about whether a monkey's leg makes a part of a tree trunk more probable/common (than the part of the tree trunk is by itself). In fact there might be a slight increase of the probability of a part of a tree trunk in the universe if there is a monkey's leg in the universe. However, surely the monkey's leg has much more positive influence on the probability of the existence of the rest of the monkey. Similarly, the part of the tree trunk is more a part of the tree in this sense than of the monkey's leg. These arguments concerned the spatial identity problem. However, for temporal identity similar arguments hold. The existence of a living organism at one point in time makes it more probable that there is a living organism (in the vicinity) at a subsequent (and preceding) point in time. If we look at structures that are not entity-like with respect to the temporal dimension we find a different situation. A part of a tree trunk at some instance of time does not make the existence of a monkey's leg at a subsequent instance much

[4]Note that in Section 5.5 we will also see that non-interpenetration of entities allows a uniquely defined notion of perception. A uniquely defined notion of perception in the case of interpenetrating entities still eludes us.

[5]We thank Eric Smith for pointing out this example.

more probable. It makes the existence of a tree at a subsequent instance much more probable. So the part of the tree trunk seems to be more a temporal instance/part of the tree than of the monkey's leg. For STPs we can easily formalise such intuitions. We required that for an entity every part of the structure, which is now a STP x_O, makes every other part more probable. A part of a STP is a STP x_b with $b \subset O$. If we require that every part of a STP makes every other part more probable then we can write that x_O is an entity if:

$$\min_{b \subset O} \frac{p_{O \setminus b}(x_{O \setminus b} | x_b)}{p_{O \setminus b}(x_{O \setminus b})} > 1. \tag{5.9}$$

This is equivalent to

$$\min_{b \subset O} \frac{p_O(x_O)}{p_{O \setminus b}(x_{O \setminus b}) p_b(x_b)} > 1. \tag{5.10}$$

If we write $\mathfrak{L}_2(O)$ for the set of all bipartitions of O we can rewrite this further as

$$\min_{\pi \in \mathfrak{L}_2(O)} \frac{p_O(x_O)}{\prod_{b \in \pi} p_b(x_b)} > 1. \tag{5.11}$$

We can interpret this form as requiring that for every possible partition $\pi \in \mathfrak{L}_2(O)$ into two parts x_{b_1}, x_{b_2} the probability of the whole STP $x_O = (x_{b_1}, x_{b_2})$ is bigger than its probability would be if the two parts were independent. To see this, note that if the two parts x_{b_1}, x_{b_2} were independent we would have

$$p_O(x_O) =: p_{b_1, b_2}(x_{b_1}, x_{b_2}) = p_{b_1}(x_{b_1}) p_{b_2}(x_{b_2}). \tag{5.12}$$

Which would give us

$$\frac{p_O(x_O)}{\prod_{b \in \pi} p_b(x_b)} = 1 \tag{5.13}$$

for this partition.

From this point of view the choice of bipartitions only seems arbitrary. For example, the existence a partition ξ into three parts such that

$$p_O(x_O) = \prod_{c \in \xi} p_c(x_c) \tag{5.14}$$

seems to suggest that the STP x_O is not an entity but instead composite of three parts. We can therefore generalise Eq. (5.11) to include all partitions $\mathfrak{L}(O)$ (see Theorem 1) of O except the unit partition $\mathbf{1}_O$ (Definition 4). Then we would say that x_O is an entity if

$$\min_{\pi \in \mathfrak{L}(O) \setminus \mathbf{1}_O} \frac{p_O(x_O)}{\prod_{b \in \pi} p_b(x_b)} > 1. \tag{5.15}$$

This measure already results in the same entities as the measure we propose.

However, in order to connect with information theory, log-likelihoods, and related literature we formally introduce the logarithm into this equation. For this we use the definition of specific local integration $\mathrm{mi}_\pi(x_O)$ of a STP x_O with respect to a partition $\pi \in \mathfrak{L}(O)$ (see Definition 49) as

$$\mathrm{mi}_\pi(x_O) := \log \frac{p_O(x_O)}{\prod_{b \in \pi} p_b(x_b)}. \tag{4.9 revisited}$$

Then according to Definition 53 we can write the complete local integration $\iota(x_O)$ of a STP as

$$\iota(x_O) := \min_{\pi \in \mathfrak{L}(O) \setminus \mathbf{1}_O} \mathrm{mi}_\pi(x_O). \qquad \text{(4.72 revisited)}$$

Finally, we can define *ι-entities* as those STPs that are completely locally integrated.

Definition 73 (ι-entity)**.** *Given a multivariate Markov chain $\{X_i\}_{i \in V}$ a STP x_O is a ι-entity if*

$$\iota(x_O) > 0. \tag{5.16}$$

The ι-entity-set $\mathfrak{E}_\iota(\{X_i\}_{i \in V})$ is then defined as follows.

Definition 74 (*ci*-entity-set)**.** *Given a multivariate Markov chain $\{X_i\}_{i \in V}$ the ι-entity-set is the entity-set*

$$\mathfrak{E}_\iota(\{X_i\}_{i \in V}) := \{x_O \in \bigcup_{A \subseteq V} \mathcal{X}_A : \iota(x_O) > 0\}. \tag{5.17}$$

Note, that due to the disintegration theorem (Theorem 22) $\mathfrak{E}_\iota(\{X_i\}_{i \in V})$ contains the same elements as the union of the refinement-free disintegration hierarchies over all trajectories:

$$\mathfrak{E}_\iota(\{X_i\}_{i \in V}) = \bigcup_{x_V \in \mathcal{X}_V} \mathfrak{D}^\blacktriangleleft(x_V). \tag{5.18}$$

5.3.5.3 Interpretations and relations

The notion of ι-entities can be interpreted in multiple ways. The introduction of the logarithm into our formalism might seem arbitrary. However, it leads to connections to other considerations especially in information theory and inference. Here we list some of these connections.

- A first consequence of introducing the logarithm is that we can now formulate the condition of Eq. (5.16) analogously to an old phrase attributed to Aristotle that "the whole is more than the sum of its parts". In our case this would need to be changed to "the log-probability of the (spatiotemporal) whole is greater than the sum of the log-probabilities of its (spatiotemporal) parts". This can easily be seen by rewriting Eq. (4.9) as:

$$\mathrm{mi}_\pi(x_O) = \log p_O(x_O) - \sum_{b \in \pi} \log p_b(x_b). \tag{5.19}$$

- Another side effect of using the logarithm is that we can interpret Eq. (5.16) in terms of the surprise value (also called information content) $-\log p_O(x_O)$ MacKay (2003) of the STP x_O and the surprise value of its parts with respect to any partition π. Rewriting Eq. (4.9) using properties of the logarithm we get:

$$\mathrm{mi}_\pi(x_O) = \sum_{b \in \pi} (-\log p_b(x_b)) - (-\log p_O(x_O)).$$

Interpreting Eq. (5.16) from this perspective we can then say that a STP is an entity if the sum of the surprise values of its parts is larger than the surprise value of the whole.

- With respect to hypothesis testing, we can view the product probability $\prod_{b\in\pi} p_b(x_b)$ with respect to partition π as the probability of x_O associated with the hypothesis that the parts x_b are stochastically independent. Let us call this hypothesis $\mathcal{H}_\pi$. Then we can write:

$$p(x_O|\mathcal{H}_\pi) := \prod_{b\in\pi} p_b(x_b). \tag{5.20}$$

Similarly, we can view the joint probability $p_O(x_O)$ as the probability of x_O under the hypothesis that the full joint probability is needed. Let us write $\mathcal{H}_\mathbf{1}$ for this hypothesis and define accordingly:

$$p(x_O|\mathcal{H}_\mathbf{1}) := p_O(x_O). \tag{5.21}$$

The occurrence of x_O is then said to provide what is called the "weight of evidence in favour of $\mathcal{H}_\mathbf{1}$" (MacKay, 2003) defined by

$$\log \frac{p(x_O|\mathcal{H}_\mathbf{1})}{p(x_O|\mathcal{H}_\pi)} > 0. \tag{5.22}$$

So in this terminology a completely locally integrated STP x_O provides evidence in favour of $\mathcal{H}_\mathbf{1}$ compared to *each* hypothesis $\mathcal{H}_\pi$, $\pi \in \mathfrak{L}(O)\backslash\mathbf{1}$ that supposes it is composite of stochastically independent parts.

- In coding theory, the Kraft-McMillan theorem (Cover and Thomas, 2006) tells us that the optimal length (in a uniquely decodable binary code) of a code word for an event x is $l(x) = -\log p(x)$ if $p(x)$ is the *true* probability of x. If the encoding is not based on the true probability of x but instead on a different probability $q(x)$ then the difference between the optimal code word length and the chosen code word length is

$$-\log q(x) - (-\log p(x)) = \log \frac{p(x)}{q(x)}. \tag{5.23}$$

Then we can interpret the specific local integration as a difference in code word lengths. Say we want to encode what occurs at the nodes/random variables indexed by O i.e. we encode the random variable X_V. We can encode every event (now a STP) x_O based on $p_O(x_O)$. Let's call this the *joint code*. Given a partition $\pi \in \mathfrak{L}(O)$ we can also encode every event x_O based on its product probability $\prod_{b\in\pi_O} p_b(x_b)$. Let's call this the *product code with respect to* π. For a particular event x_O the difference of the code word lengths between the joint code and the product code with respect to π is then just the specific local integration with respect to π.

Complete local integration then requires that the joint code code word is shorter than all possible product code code words. This means there is no partition with respect to which the product code for the STP x_O has a shorter code word than the joint code. So entities are STPs that are shorter to encode with the joint code than a product code.

- We can relate our measure of identity to other measures in information theory. For this we note that the expectation value of specific local integration with respect to a partition π is the multi-information $\mathcal{I}_\pi(X_O)$ McGill (1954); Amari (2001) with respect to π, i.e.

$$\mathcal{I}_\pi(X_O) := \sum_{x_O \in \mathcal{X}_O} p_O(x_O) \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)} \tag{5.24}$$

$$= \sum_{x_O \in \mathcal{X}_O} p_O(x_O)\, \mathrm{mi}_\pi(x_O). \tag{5.25}$$

  The multi-information plays a role in measures of complexity and information integration (Ay, 2015). The generalisation from bipartitions to arbitrary partitions is applied to expectation values similar to the multi-information above in Tononi (2004). The relations of our localised measure (in the sense of Lizier (2012)) to multi-information and information integration measures also motivates the name specific *local integration*. Relations to these measures will be studied further in the future. Here we note that these are not suited for measuring identity of STPs since they are properties of the random variables X_O and not the values x_O.

- Using the disintegration theorem (Theorem 22) results in yet another point of view. The theorem states that for each trajectory $x_V \in \mathcal{X}_V$ of a multivariate Markov chain the refinement-free disintegration hierarchy only contains completely integrated STPs i.e. it only contains ι-entities. It also contains all ι-entities that occur in that trajectory. The disintegration hierarchy is obtained by sorting the partitions $\pi \in \mathfrak{L}(V)$ of the trajectory x_V according to increasing specific local integration $\mathrm{mi}_\pi(x_V)$ of x_V. This results in the disintegration levels $\mathfrak{D}_i(x_V)$ with $\mathfrak{D}_1(x_V)$ containing the partitions with the least specific local integration. To get to the refinement-free version of the disintegration hierarchy, we remove all partitions from each level $\mathfrak{D}_i(x_V)$ that either have a refinement at that level or have a refinement at a lower level $\mathfrak{D}_j(x_V)$ with $j < i$. A partition in the refinement-free disintegration hierarchy is always a minimal/finest partition (Definition 3) reaching such a low specific local integration.

  Each ι-entity is then a block x_c with $c \in \pi$ of a partition $\pi \in \mathfrak{D}^\blacktriangleleft(x_V)$ for some trajectory $x_V \in \mathcal{X}_V$ of the multivariate Markov chain.

  Let us recruit the interpretation from coding theory above. If we want to find the optimal encoding for the entire multivariate Markov chain $\{X_i\}_{i\in V}$ this means finding the optimal encoding for the random variable X_V whose values are the trajectories $x_V \in \mathcal{X}_V$. The optimal code has the code word lengths $-\log p_V(x_V)$ for each trajectory x_V. The partitions in the lowest level $\mathfrak{D}_1^\blacktriangleleft(x_V)$ in the refinement-free disintegration hierarchy for x_V have minimal specific local integration i.e.

$$\mathrm{mi}_\pi(x_V) = \log \frac{p_V(x_V)}{\prod_{c\in\pi} p_c(x_c)} \tag{5.26}$$

is minimal among all partitions. At the same time these partitions are the finest partitions that achieve this low specific local integration. This implies on the one hand that the code word lengths of the product codes associated to these partitions are the shortest possible for x_V among all partitions. On the other hand these partitions split up the trajectory in as many parts as possible while generating these shortest code words. In this combined sense the partitions in $\mathfrak{D}_1^{\blacktriangleleft}(x_V)$ generate the "best" product codes for the particular trajectory x_V.

Note that the *expected code word length* of the product code:

$$\sum_{x_V \in \mathcal{X}_V} p_V(x_V)(-\log \prod_{c \in \pi} p_c(x_c)) \tag{5.27}$$

which is the more important measure for encoding in general, might not be short at all. The product codes based on partitions in $\mathfrak{D}_1^{\blacktriangleleft}(x_V)$ are specifically adapted to assign a short code word to x_V i.e. to a single trajectory or story of this system. They are constructed/forced to describe x_V as a composition of stochastically independent parts. More precisely they are constructed in the way that would be optimal for stochastically independent parts. The parts themselves are chosen to minimise $\mathrm{mi}_\pi(x_V)$ for x_V.

Nonetheless, the product codes exist (they can be generated using Huffman coding or arithmetic coding Cover and Thomas (2006) based on the product probability) and are uniquely decodable. What would they be useful for? Say for some reason the trajectory x_V is more important than any other and that we want to "tell its story" as a story of as many as possible (stochastically) independent parts (that are maybe not really stochastically independent) i.e. we wanted to encode the trajectory *as if it were* a combination of as many as possible stochastically independent parts/events. And because x_V is more important than all other trajectories we wanted the code word for x_V to be the shortest possible. Then we would use the product codes of partitions in the refinement-free disintegration hierarchy because those combine exactly these two conditions. The pseudo-stochastically-independent parts would then be the blocks of these partitions which according to the disintegration theorem are exactly the ι-entities occurring in x_V.

On a very speculative note we mention that the trajectory/history that we (real living humans) live in is more important to us than all other possible trajectories of our universe (if there are any). What happens in this trajectory needs to be communicated more often than what happens in counterfactual trajectories. Furthermore a good reason to think of a system as composite of as many parts as possible is that this reduces the number of parameters that need to be learned which in turn improves the learning speed (see e.g. Kolchinsky and Rocha, 2011). So the entities that mankind has partitioned our history into might somehow serve a purpose related to the product codes generated form partitions of the refinement-free disintegration hierarchy of our universe.

Recall that this kind of product code is not the optimal code in general (which would be the one with shortest expected code word length). It is possibly more of a naive code that does not require deep understanding of the dynamical system but instead can be learned fast and works. The language of physics for example might be more optimal in the sense of shortest expected code word lengths reflecting a desire to communicate efficiently about all counterfactual possibilities as well.

This concludes the motivation of our proposal to use completely locally integrated patterns as entities in multivariate Markov chains. Next we will present definitions of actions and perceptions.

5.4 Entity action

Here we define a concept of actions for a given entity set (Definition 65) in a multivariate Markov chain. First we discuss some challenges that arise when trying to define actions within such rigidly defined systems (Section 5.4.1). Then we motivate our approach to actions (Section 5.4.2). In Section 5.4.3, we finally present the formal definition of actions for entities in multivariate Markov chains. Conceptually, this section is loosely related to our own publication Biehl and Polani (2015) but the formal setting is different. The formal definition of actions in Section 5.4.3 to our knowledge is the first of its kind. This is to say that it is the first formal definition of actions that is applicable to individuals/entities within multivariate Markov chains (including dynamical system, cellular automata etc.).

In order to avoid confusion we will refer to individuals whenever we speak of entities that are not necessarily elements of an entity set in the technical sense of Definition 65. For example, we speak of animals as individuals that can perform actions. This does not imply that there is another notion of "individual" which needs to be defined. We argued in Section 5.3 that in the context of multivariate Markov chains such individuals correspond to entities. Outside of multivariate Markov chains we have not made such arguments and therefore use the term individual here.

5.4.1 Contrast to more common conceptions

Paraphrasing Wilson and Shpall (2012) only slightly, what distinguishes actions among events or occurrences is that they do not merely happen to individuals but rather that they are *made to happen by* the individuals.

This is problematic in our setting where STPs (as entities) take the role of individuals. What "happens" in a multivariate Markov chain are the trajectories and the STPs occurring in them. The Markov chain's dynamics are determined by its mechanisms $p_{j,t}$ with $j \in J, t \in T$. These in turn determine (possibly stochastically) what is going to happen anywhere within the chain. All mechanisms at all time-steps are fixed by the definition of the Markov chain and then cannot be altered anymore. If it is desired that mechanisms change over time then this must be decided when defining the Markov chain.

Since the occurrence of STPs is an effect of these fixed mechanisms the STP cannot "make anything occur" within the chain. Just like the occurrence of any STP up to time t is a consequence of the mechanisms so are the occurrences of STPs in the future of t. More formally, given any STP (be it an entity or not) x_A, its morph (see Definition 26) $p_{V\setminus A|A}(X_{V\setminus A}|x_A)$ is the probability distribution over the rest possible states $\mathcal{X}_{V\setminus A}$ of the multivariate Markov chain given that x_A occurs. By definition this morph is determined uniquely by the mechanisms of the chain. This means whatever "happens" beyond the STP x_A is already determined when the Markov chain is defined.

Therefore, it is impossible that a multivariate Markov chain contains an STP or entity that can make something happen beyond what happens anyway due to the mechanisms. This means we have to explain and define actions in a different way.

Before we go on we should note that many accounts of actions require the actions to be in the interest of some goal or to serve some purpose (Wilson and Shpall, 2012). In accordance with our working definition we view goal-directedness as a separate phenomenon and will not follow the practice of requiring such for actions themselves. In our case an entity with actions will be considered goal-directed if its actions are goal-directed in some sense.

After these comments on what we cannot do and what we choose not to do we will now motivate our own approach. First, we give some background and observations about actions that motivate our definition. Then we present the main ideas behind it and finally state the definition.

5.4.2 Background to our concept of actions

We can make two observations about the common (human) usage of the term action. The first is that events called actions are usually attributed to a limited or bounded region or part of the universe e.g. the body of a living organism or sometimes just its brain if it has one. These parts usually contain mechanisms or configurations of matter that are either

- not directly observable to a human observer e.g. hidden in an opaque container, or
- not well understood by the human observer, or
- both.

These factors inevitably lead to unpredictability of such events. In other words, events that are attributed to well understood and therefore predictable mechanisms, e.g. sunrises, are not considered actions.

Let us consider the above more closely. Historically, actions (agency) have been attributed to more things than just animals or living systems (or robots). An example of this is the attribution of natural phenomena like thunder and lightning to divine interventions in Rome in 50 BCE which was criticised by Lucretius (2007). Lightning in particular was often seen as a goal-directed action by the god Zeus; the goal being to punish humans. Later, in the 19th century Nietzsche (1892, pp.26) criticised the separation of "the lightning" (der

Blitz) as a subject and the flashing light (das Leuchten) as its action (Thun). It is notable that as science progressed it was able to explain more and more phenomena without divine (or any other) interventions and without reference to any actions (or goal-directedness) at all [6]. Also note that, the mechanics behind lightning and thunder were difficult to understand before technological and scientific advances and still are difficult to observe as they are due to electrodynamics (and happen in places that are hard to access).

Nowadays we have a mechanistic account and events like thunder and lightning are hardly considered more special than an apple falling to the ground because of gravity. Actions do remain to be attributed to animals of course. Most prominently to humans and their nervous systems. These systems coincide with the most complex known parts of the known universe i.e. those parts that are extremely hard to understand . From our point of view this is not a coincidence. It is the complexity and opacity of these mechanisms that make us attribute actions to them. If we would have the sensory and computational capacity to watch and keep track of the dynamics of entire brains, we believe that it would look to us again like an apple falling to the ground. From this point of view actions are not, beyond their possibly complex and unobserved origin, special events but may *appear* as such to observers that lack the sensory and computational capacity to resolve or understand them. This suggests that for actions to occur within a system there needs to be both observers and corresponding mechanisms that exceed the capacity of those observers to resolve them i.e. see them as mere consequences of the dynamical law. Note that the observers might themselves be such opaque mechanisms for other observers and for themselves.

There may be a possibility to define actions in a fundamental way without the need to define observers first. Say there are events within the universe which are as a matter of principle not distinguished by *any* observer. Then events occurring as a consequence of these events will be inexplicable for any observer. So these events will appear to be actions in general. This is the route we take below. Note that this approach remains compatible with an observer-dependent notion of actions. The "fundamental" actions are apparent actions for every possible observer while other events are actions for some observers and "plainly" predictable events for others.

What we have ignored in this discussion up to now is the role of randomness. True randomness (in the sense of stochastic independence of the event from any other event in the universe), if it exists in a universe, can never be explained, predicted, or understood. Combined with our reasoning above this suggests that all random events are actions and even fundamental actions in the sense that no observer could possibly resolve the different events that lead to the random events just because there are no different events that lead to a random event. The random event happens independently of everything else. This is also the reason why we would not like to see random events as actions. They are not the result of some indistinguishable but in-system events. Even

[6]The term "action" in the "principle of least action" plays a major role in physics. However, this principle is used to determine trajectories of dynamical systems and has no relation to possible actions performed by parts of the system. It is therefore ignored in the discussion here.

an all-observing being external to the universe could not predict them from the internals.

One way to avoid random events being mistaken for actions would then be to require that external observers, which are not limited by the restrictions on observability for internal observers, can predict the action events from other events that are internally unobservable. Note that this could be seen as the adaptation to our setting of the widely accepted view that actions are initiated by the agent (e.g. Schlosser, 2015). In our definition of actions we do not explicitly require this. Instead, the burden to avoid the random events from being mistaken for actions is put on the choice of the entity set. In our conception of actions, actions can only be performed by entities, more precisely they can only occur as parts of larger entities. Intuitively entities are spatiotemporal-patterns whose parts are in some way connected to each other. Random events (stochastically independent from all other events) are therefore not expected to be parts of entities. If the entities in a given entity-set do not contain parts that are random events an explicit requirement of predictability is not needed. Note that, according to trivial definitions of entities (like the unrestricted mereological composition of Section 5.3.5) random events may be parts of larger entities. In that case we expect non-intuitive consequences anyway.

5.4.3 Definition of actions for entities

When we want to define actions for entities the first issue we run into is that entities are already fixed STPs. They may or may not have "acted" within a trajectory that they occur in but once we have the entity its "story" is fixed. In order to define actions we therefore look at the sequence of time-slices of an entity and investigate what "could have happened". In the end, whenever there are counterfactual entities that could have taken the place of the entity without changes in the rest of the system and then went on a different path we will say that an action occurred.

In more detail, first note that an action always requires the possibility of an alternative action. However, as argued before, a single entity x_A occurring in trajectory x_V has no alternative options since the trajectory determines everything. Therefore for an action of entity x_A in trajectory x_V we require the existence of alternative/counterfactual entity y_B in another trajectory y_V. For an action to occur at time t

- the entities x_A and y_B must occupy the same random variables at t i.e. $A_t = B_t$,
- the time-slices x_{v_t} and y_{V_t} at t of the two trajectories must coincide everywhere apart from the random variables that are occupied by the entities i.e. $x_{V_t \setminus A_t} = y_{V_t \setminus A_t}$.

It is then impossible that any observer that is in the "environment" $x_{V_t \setminus A_t}$ of the entities can distinguish the entities because the states of all such observers are identical in both trajectories. For an action to occur the two identities must

then differ at time $t+1$. We define the environment of a STP here for further use.

Definition 75 (Environment of an STP)**.** *Let $\{X_i\}_{i\in V}$ be a multivariate Markov chain with $V = J \times T$ and let x_A be a STP. Then the environment of x_A at time t is the spatial pattern $x_{V_t \setminus A_t}$.*

As mentioned before we do not require that the difference at time $t+1$ is predictable from the entities during the interval. Such relations between parts of entities, if desired, must be imposed by the choice of the entity set.

This construction may lead to the following question. According to this definition actions rely on counterfactual trajectories. However, actions as commonly understood occur all the time within the single history/trajectory that we are experiencing. Since nobody has ever experienced two alternative trajectories of our universe the question is how can this concept play a role in our conception of the world? The answer to this is that the existence of actions as we defined them will force conceptions or models of the world to incorporate them. Let us assume that humans model the universe they exist in to some degree. According to the indistinguishability requirement whenever there is an action by another entity this model will lack the data to distinguish which act will occur. In such situations it should be prepared for both acts, i.e. it should model both acts. So the counterfactual trajectory of the universe becomes relevant for individuals modelling their environment/world.

Another question may concern the effect of actions according to this definition. We have not required that the actions i.e. the different time-slices of the counterfactual entities at time $t+1$ are distinguished by any "observer". While such requirements may be possible we make no such requirement here. Our definition of actions is deliberately weak. As mentioned before it is the entity set that we see as selective. In the future further notions of actions will be investigated.

We now state the definition of an action of an entity at a time t in a particular trajectory formally.

Definition 76 (Action and co-action of an entity)**.** *Let $\{X_i\}_{i\in V}$ be a multivariate Markov chain with $V = J \times T$ and let $x_V \in \mathcal{X}_V$ with $p_V(x_V) > 0$. Also let x_A be an entity with non-empty time-slices at $t, t+1$. Then x_A performs an action $x_{A_{t+1}}$ at time t in trajectory x_V if there exists an entity y_B with non-empty time-slices at $t, t+1$ such that*

(i) y_B occurs in $y_V \neq x_V$ with $p_V(y_V) > 0$,
(ii) at t the entities x_A and y_B occupy the same random variables: $B_t = A_t$,
(iii) at t the trajectories x_V and y_V are otherwise identical: $x_{V_t \setminus A_t} = y_{V_t \setminus A_t}$,
(iv) at $t+1$ the entities are different: $x_{A_{t+1}} \neq y_{B_{t+1}}$.

We also call y_B a co-action entity, y_V a co-action trajectory, and $y_{B_{t+1}}$ a co-action.

Remark:

- Note that all requirements are symmetric. Therefore, if x_A performs an action $x_{A_{t+1}}$ at time t in trajectory x_V then also y_B performs an action $y_{B_{t+1}}$ at time t in trajectory y_V. This motivates our terminology of *co*-actions.

- The notion of co-action entities can easily be extended to more than one co-action entity. We only have to make sure that all entities in a set of co-action entities are mutually different at $t+1$.
- A further requirement that we could make here would be that y_B does not occur in x_V. This is not excluded in this definition. At time t the two entities can in principle be equal $x_{A_t} = y_{B_t}$. At $t+1$ we could have $A_{t+1} \cap B_{t+1} = \varnothing$ so that even if $x_{A_{t+1}} \neq y_{B_{t+1}}$ we can have $x_{B_{t+1}} = y_{B_{t+1}}$. This requires that entities can be identical at some time t and then different at some time $t+1$. We do not exclude this possibility here. It is an interesting question for further research at what level such situations should be prevented (if it should be prevented). It could be introduced as an axiom for entity sets which corresponds to prohibiting interpenetration of entities. However, it could also be a selective criterion for specific dynamics of the multivariate Markov chain. One could imagine that there is a set of dynamics obeying a certain conservation law that prevents interpenetration. The notion of ι-entities does not prevent interpenetration as we see in Section 6.5.
- It is easy to generalise the definition of actions to situations where x_A and y_B must occupy the same variables for an interval of time $[t-m:t]$ before the action. In that case, the environment $x_{V_{[t-m:t]}} \setminus A_{[t-m:t]}$ must also be identical during this interval.

The condition that the two acting entities differ at time $t+1$ can be fulfilled in two ways. The entities can differ in this time-slice in value or in extent.

Definition 77 (Value and extent actions)**.** *Let $\{X_i\}_{i \in V}$ be a multivariate Markov chain with $V = J \times T$. If x_A performs an action $x_{A_{t+1}}$ at time t in trajectory x_V and $y_{B_{t+1}}$ is its co-action we can distinguish two special cases of actions:*

(i) if the actions differ in extent i.e. we have

$$A_{t+1} \neq B_{t+1} \tag{5.28}$$

then we call these actions extent actions.

(ii) if the actions differ only in value i.e. we have

$$A_{t+1} = B_{t+1} \tag{5.29}$$

so that

$$x_{A_{t+1}} \neq y_{A_{t+1}}. \tag{5.30}$$

then we call these actions value actions.

Remarks:

- Value actions are a particularly weak notion of action in some sense. Since we define the action only as a difference to the co-action. The environment (or the entire future) may stay unaffected by such an "action". The entire morph can be identical for such actions i.e.

$$p_{V \setminus A_{t+1}}(X_{V \setminus A_{t+1}} | x_{A_{t+1}}) = p_{V \setminus A_{t+1}}(X_{V \setminus A_{t+1}} | y_{A_{t+1}}). \tag{5.31}$$

Note that the extent actions always have an effect since they change the random variables that are part of the entity. Formally, the morphs of two different extent actions are always different because they range over different variables. Stronger definitions of value and extent actions which require for example that the environments change or that their morphs differ in particular respects are also possible and may have their own merits. However, an investigation of different definitions is beyond the scope of this thesis.

The difference between value actions and extent actions is made possible due to our definition of entities as STPs. We have argued in Section 5.3.3 that entities should vary counterfactually in value and extent. An intriguing question for the future is whether the capabilities of agents to act both in value and extent are truly superior to agents that only act in value. With regard to the theory of computation in distributed systems by Lizier et al. (2014) one can also ask whether there are computational advantages to either. As we will see in Section 5.6 probabilistic and information theoretic expressions are easy to formulate for actions in value only. However, for actions in extent this has not been done yet.

5.5 Entity perception

In this section we formally define perception for entities in multivariate Markov chains. We make no distinction here between perception, experience, and sensory input. In the tradition of modelling agent-environment systems using dynamical systems or their probabilistic generalisations stochastic processes we define perception as *all effects* that the environment has on an individual/agent (Beer, 1995). In contrast to previous work along this line the individuals are not modelled as a dynamical system coupled to the environment (Beer, 1995; Der et al., 1999; Ay et al., 2012) or a stochastic process interaction with another one (Klyubin et al., 2004; Lungarella et al., 2005; Bertschinger et al., 2008; Seth, 2010; Ay et al., 2012). In our case the individuals are entities i.e. special kinds of STPs. In order to define perception we therefore have to capture all effects of the environment on *entities* or if we focus on perception of a single entity all effects on that entity.

Thinking about this we run into a similar problem as with the actions. An entity is already a fixed STP that contains all influence that it may have been subjected to. It is in this sense the *result* of influence (or no influences) from its surroundings. In order to investigate these influences we therefore have to deconstruct the entity and see how it was "formed" by external influences / perceptions time-slice by time-slice.

The idea here is to use the same (or a similar) construction as in the extraction of sensor-values for the extended perception-action loop in Section 3.3.6. As we have seen there this construction of sensor-values captured all influences of the environment process on the agent process. This was established by showing that the dependence on the environment can be replaced by the dependence on the sensor-values without changing the agent or environment

processes. So whatever influence the environment process has on the agent process, this influence is contained in the sensor-values. Underlying this construction is the classification of the environment into classes that have identical influence on the transition of the agent process from on time-step to the next. More precisely, the sensor-values of the extended perception-action loop are constructed as equivalence classes of environments with respect to the conditional probability distributions $p_{M_{t+1}}(.|m_t, e_t) : \mathcal{M}_{t+1} \to [0,1]$. We defined that two environments $\hat{e}_t, \bar{e}_t \in \mathcal{E}_t$ at time t are produce the same perception / sensor value if they induce the same conditional probability distribution over the agent's next time-step:

$$\hat{e}_t \equiv_{\epsilon_t} \bar{e}_t \Leftrightarrow \forall m_{t+1} \in \mathcal{M}_{t+1}, m_t \in \mathcal{M}_t : p_{M_{t+1}}(m_{t+1}|m_t, \hat{e}_t) = p_{M_{t+1}}(m_{t+1}|m_t, \bar{e}_t). \tag{3.71 revisited}$$

Here ϵ_t is the partition induced by this equivalence relation. In this section we are interested in defining the influence of the environment on an entity. This will require a generalisation of Eq. (3.71) which involves some subtleties.

Before we present the generalisation let us look at a simple example of the standard construction. This will lead to a better intuition for our concept of perception.

5.5.1 Example of perception in the perception-action loop

Say we have a binary agent process $\mathcal{M}_t = \{m_t^1, m_t^2\}$ and a ternary environment process $\mathcal{E}_t = \{e_t^1, e_t^2, e_t^3\}$. For a given fixed value $m_t \in \mathcal{M}_t$ of the agent memory at t each environment value $e_t \in \mathcal{E}_t$ then has an associated conditional probability distribution $p_{M_{t+1}}(.|m_t, e_t) : \mathcal{M}_{t+1} \to [0,1]$ of the form:

	m_{t+1}^1	m_{t+1}^2
$p_{M_{t+1}}(.\|e_t^1, m_t)$	q	$1-q$
$p_{M_{t+1}}(.\|e_t^2, m_t)$	r	$1-r$
$p_{M_{t+1}}(.\|e_t^3, m_t)$	s	$1-s$

(5.32)

where $q, r, s \in [0,1]$. First, assume $q = r = s$. We then have for all $m_{t+1} \in \mathcal{M}_{t+1}$

$$p_{M_{t+1}}(m_{t+1}|m_t, e_t^1) = p_{M_{t+1}}(m_{t+1}|m_t, e_t^2) = p_{M_{t+1}}(m_{t+1}|m_t, e_t^3). \tag{5.33}$$

Since we are eventually interested in fixed realisations of entities (and not in random variables), we here drop the requirement of Eq. (3.71) that Eq. (5.33) needs to hold for all $\bar{m}_t \in \mathcal{M}_t$ and consider the equivalence classes that are generated for the *specific* $m_t \in \mathcal{M}_t$. We then get

$$e_t^1 \equiv_{\epsilon_t} e_t^2 \equiv_{\epsilon_t} e_t^3. \tag{5.34}$$

This means there is only a single block in ϵ_t i.e. $\epsilon_t = \{\{e_t^1, e_t^2, e_t^3\}\}$. So all environments have the same influence on the next agent state m_{t+1} given m_t. Or, equivalently, given m_t, no differences in the environment make a difference to m_{t+1}. Since we equate influence with perception here, we interpret this as saying that no perception occurs in the transitions from m_t into $\mathcal{M}_{t+1}$.

Second, assume that $q = r \neq s$. Then $\epsilon_t = \{\{e_t^1, e_t^2\}, \{e_t^3\}\}$ containing two blocks. This means that e_t^1 and e_t^2 influence the transition from m_t into $\mathcal{M}_{t+1}$ in the same way while e_t^3 has a different influence. We then say that in the transition from m_t into $\mathcal{M}_{t+1}$ there are two perceptions/sensor values corresponding to the two blocks of ϵ_t. Note that while we have perception in this transition the perception is not perfect. It cannot distinguish between e_t^1 and e_t^2.

Third, assume that $q \neq r \neq s \neq q$. Then $\epsilon_t = \{\{e_t^1\}, \{e_t^2\}, \{e_t^3\}\}$ containing three blocks and fully resolving the environment. This means each environment influences the transition from m_t into $\mathcal{M}_{t+1}$ differently. We then have three different perceptions in the transition from m_t into $\mathcal{M}_{t+1}$ resolving the environment states fully.

So our notion of perception employs differences in the influence of environments on transitions from a value m_t to its possible successors $\mathcal{M}_{t+1}$ to classify the environments. For entities we will try to use the same approach with some necessary generalisations. for this it is also helpful to note the following.

Instead of only considering the next time-step note that we can also use the next two time-steps (or any number $r \in [t+1 : n-1]$ of next time-steps[7], where $n-1$ is the last time-step in T) in the same way to partition the environment. For example if we consider the next two time-steps $t+1, t+2$ we can define the equivalence classes of environments via:

$$\begin{aligned} &\hat{e}_t \equiv_{\epsilon_t^2} \bar{e}_t \\ &\quad \Leftrightarrow \forall m_{t+1} \in \mathcal{M}_{t+1}, m_t \in \mathcal{M}_t : \\ &\quad\quad p_{M_{t+1},M_{t+2}}(m_{t+1}, m_{t+2}|m_t, \hat{e}_t) = p_{M_{t+1},M_{t+2}}(m_{t+1}, m_{t+2}|m_t, \bar{e}_t). \end{aligned} \tag{5.35}$$

The resulting partition ϵ_t^2 is then a refinement of the partition ϵ_t. To see this note that

$$p_{M_{t+1}}(m_{t+1}|m_t, e_t) = \sum_{m_{t+2} \in \mathcal{M}_{t+2}} p_{M_{t+1},M_{t+2}}(m_{t+1}, m_{t+2}|m_t, e_t) \tag{5.36}$$

such that all environments $\hat{e}_t, \bar{e}_t \in \mathcal{E}_t$ that are in distinct blocks of ϵ_t i.e. those with

$$p_{M_{t+1}}(m_{t+1}|m_t, \hat{e}_t) \neq p_{M_{t+1}}(m_{t+1}|m_t, \bar{e}_t) \tag{5.37}$$

are also in distinct blocks of ϵ_t^2 because Eqs. (5.36) and (5.37) imply

$$p_{M_{t+1},M_{t+2}}(m_{t+1}, m_{t+2}|m_t, \hat{e}_t) \neq p_{M_{t+1},M_{t+2}}(m_{t+1}, m_{t+2}|m_t, \bar{e}_t). \tag{5.38}$$

The more time-steps into the future we consider the finer the induced partition of the environment. Conversely, the partition of the environment obtained by considering only one next time-step is a coarsening of those obtained by considering more time-steps.

[7] In fact the same construction can be used with any subset of the future times $[t+1 : n-1]$. We will not pursue this generalisation further in this thesis.

5.5.2 Steps to get perception for entities

In order to get a notion of perception for entities that is similar to the notion of perception based on influence that we used for the perception-action loop the intuition is then simply

- "take an entity at time t" (analogous to m_t),
- get the "possible next time-slices of this entity" at $t+1$ (analogous to $\mathcal{M}_{t+1}$ above),
- obtain "the" conditional probability distribution over these "next time-slices" given the current one and the environment (analogous to $p_{M_{t+1}}(.|m_t, e_t)$)
- classify the environments according to their influence on the transitions to these "next time-slices".

There are, however, multiple problems which complicate the formal definition of the required notions. Some more obvious ones are listed next. We discuss them and more subtle ones in more detail in the subsequent sections.

1. An entity at time t is either a spatial entity (no temporal extension) or it is a time-slice of an entity. If it is a spatial entity then it has no "next-time slice". If it is a time-slice we are not "taking" the entire entity. If we take the entire entity then it comes with possibly long past and future extension. It can also have a past and an empty next time-slice.
2. Related to the previous point is another problem. Assuming an entity that has future extension i.e. the next-time slice is not empty, then there are no other "possible next time-slices of the entity". As mentioned before the entity is defined in its entirety. Accordingly the next time-slice of an entity (not only if it is empty) is uniquely defined just like its entire future.
3. Assuming we have obtained some "next time-slices" in a reasonable way, these may not be mutually exclusive and exhaustive unlike the values $m_{t+1} \in \mathcal{M}_{t+1}$. This means multiple next time-slices can occur together (if they are not mutually exclusive) or none of the next time-slices occurs (if they are not exhaustive). This makes the construction of the conditional probability distribution complicated.

It turns out that the steps to get a generalisation of perception for entities are more clearly presented in a slightly different order. The overarching goal remains the construction of a conditional probability distribution that generalises $p_{M_{t+1}}(.|m_t, e_t) : \mathcal{M}_{t+1} \to [0, 1]$. The steps we take in the next sections are then:

- Define entities with identical pasts up to t as analogues of "an entity at time t". These are the *co-perception entities*.
- Define the entire futures of the co-perception entities as the proto-analogues of "possible next time-slices of the entity" and only later focus on the actual next time-slices of these entities via the "branching partition".

- Devise a way to deal with the problems of non-exhaustion on the level of the entire futures of entities. To define the conditional probability distribution we need an exhaustive set of possible outcomes/futures since the sum over the possible outcomes must equal one.
- Restrict the environments that can be classified by perception to those that can co-occur with the entities.
- Deal with the problem of mutual exclusion of entity futures, which can be done by a further assumption of non-interpenetration of entities. We will see that, to define the conditional probability distribution, we need a mutually-exclusive set of possible outcomes/futures.
- Partition all co-perception entities into blocks (called branches) of entities with identical next time-slices since co-perception entities may differ at even later times only. This partition is called the branching-partition. The final conditional probability distribution over "next time-slices" will then be over the branches of his partition.
- The environments are then classified according to the conditional probability distribution over the branches.

5.5.3 Co-perception entities

We first discuss the problem that entities have unique next time-slices and therefore the set of "possible next time-slices" only contains a single time-slice. This will be resolved by using the "co-perception entities" in order to provide a set of possible next time-slices. These do not come from the same entity but from the co-perception entities. Here we motivate and discuss these entities.

First note that any part of an entity x_A (which is a STP) can also be a part of another entity y_B. This means that $A \cap B \neq \varnothing$ and $x_{A\cap B} = y_{A\cap B}$. Therefore we can also have entities x_A, y_B that are identical at some time t, i.e.

$$x_{A_t} = y_{B_t}. \tag{5.39}$$

These can in general have different next time-slices. The next time-slices of all entities that are equal to x_A at time t (where x_{A_t} is not empty) are then a first candidate for the conditional probability distribution to range over.

Note however that these entities can also have different pasts. Since we want to define the perception of a single entity we therefore only consider entities that are identical up to some time t, i.e.

$$x_{A_{\leq t}} = y_{B_{\leq t}}. \tag{5.40}$$

The set of entities with identical pasts up to time t can be interpreted as the set of entities that are the most like x_A up to t. These are *different* entities but they only differ in the future. Their futures (including their next time-slices) are therefore a close analogue to the "possible next time-slices of the entity". To make sure however that the entities have a next time-slice we also require

that they have *non-empty* next time-slice. These requirements together define the notion of the *co-perception entities of an entity x_A at time t*. These are entities that also perceive something (maybe the same thing) at t (in their trajectories) if x_A perceives something at t.

Definition 78 (Co-perception entities of an entity at t)**.** *Let $\{X_i\}_{i\in V}$ be a multivariate Markov chain with $V = J \times T$ and entity set $\mathfrak{E}$. Let $x_A \in \mathfrak{E}$ be an entity with non-empty time-slices at t and $t+1$. The set of co-perception entities $\mathfrak{S}(x_A, t)$ of entity x_A at t is the set of entities with non-empty time-slices at t and $t+1$, and that are identical up to t:*

$$\mathfrak{S}(x_A, t) := \{y_B \in \mathfrak{E} : B_t, B_{t+1} \neq \varnothing, y_{B_{\leq t}} = x_{A_{\leq t}}\}. \tag{5.41}$$

As mentioned before the time-slices at $t+1$ of the co-perception entities $\mathfrak{S}(x_A, t)$ provide an analogue of the "possible next time-slices of the entity x_A". The next step would then be to define a conditional probability distribution over this set given the past $x_{A_{\leq t}}$ of the entity and any environment $x_{V_t \setminus A_t}$ at t. Instead of only looking at the next time-slices at $t+1$ we can also consider the whole future $t <= [t+1 : n-1]$ ($n-1$ is the last time-step in T) of the co-perception entities. As we have indicated in Section 5.5.1 if we define the conditional probability distribution over the whole futures we will obtain a finer classification of the environments. We can then still refocus on the next time-slice afterwards. This general viewpoint is also more suitable for the formal development of the theory.

In order to discuss the associated problems we consider the special case of a set of co-perception entities that contains only two entities quite thoroughly. So assume that there are only two co-perception entities (including x_A itself) i.e. $\mathfrak{S}(x_A, t) = \{x_A, y_B\}$.

Apart from the conditions on co-perception entities (Definition 78) the entities in $\mathfrak{S}(x_A, t)$ are arbitrary STPs since we are trying to define perception for arbitrary entity sets.

Still we can note that since $\{x_A, y_B\}$ forms a set the two entities are not equal

$$x_A \neq y_B \tag{5.42}$$

and since $x_{A_{\leq t}} = y_{B_{\leq t}}$ (due to Definition 78) we then know that

$$x_{A_{t<}} \neq y_{B_{t<}}. \tag{5.43}$$

But we do not know at which time-slices they differ. For example they could be equal at $t+1$ or any other particular future time-step $t+r$ with $r \in [1, n-1-t]$ (at the last time step $n-1$ of the multivariate Markov chain there is no perception since there is no future). For the next time-slices $x_{A_{t+1}}$ and $y_{B_{t+1}}$ of x_A and y_B we have the extra condition that they are non-empty, i.e.

$$A_{t+1} \cap V_{t+1} \neq \varnothing \tag{5.44}$$

and

$$B_{t+1} \cap V_{t+1} \neq \varnothing. \tag{5.45}$$

Apart from this, the entities in $\mathfrak{S}(x_A, t)$ are completely arbitrary. For the following it is important to keep the possible relations between the time-slices of entities in $\mathfrak{S}(x_A, t)$ in mind. We therefore take a look at these possible relations as well as their implications for the co-occurrence (i.e. the joint probabilities) of the time-slices. The time-slices can occupy the same random variables and have the same values (i.e. be identical), occupy the same random variables and have different values, occupy partly the same random variables and have the same values at the random variables in the intersection, occupy partly the same random variables and have the different values at the random variables in the intersection, and occupy only different random variables. Formally, for the time-slices $x_{A_{t+r}}$ and $y_{B_{t+r}}$ of x_A and y_B we can have the following situations:

1. $A_{t+r} = B_{t+r}$ and $x_{A_{t+r}} = y_{B_{t+r}}$,
2. $A_{t+r} = B_{t+r}$ and $x_{A_{t+r}} \neq y_{B_{t+r}}$,
3. $A_{t+r} \neq B_{t+r}$, $A_{t+r} \cap B_{t+r} \neq \varnothing$ and $x_{A_{t+r} \cap B_{t+r}} = y_{A_{t+r} \cap B_{t+r}}$,
4. $A_{t+r} \neq B_{t+r}$, $A_{t+r} \cap B_{t+r} \neq \varnothing$ and $x_{A_{t+r} \cap B_{t+r}} \neq y_{A_{t+r} \cap B_{t+r}}$,
5. $A_{t+r} \cap B_{t+r} = \varnothing$.

So in general we have to write the probability that both of the time-slices occur (given[8] an arbitrary environment $x_{V_t \setminus A_t}$ and the identical past $x_{A_{\leq t}}$) as:

$$Pr(X_{A_{t+r}} = x_{A_{t+r}}, X_{B_{t+r}} = y_{B_{t+r}} | x_{V_t \setminus A_t}, x_{A_{\leq t}}) \tag{5.46}$$

$$\begin{aligned} =& Pr(X_{A_{t+r} \setminus B_{t+r}} = x_{A_{t+r} \setminus B_{t+r}}, X_{A_{t+r} \cap B_{t+r}} = x_{A_{t+r} \cap B_{t+r}}, \\ & X_{A_{t+r} \cap B_{t+r}} = y_{A_{t+r} \cap B_{t+r}}, X_{B_{t+r} \setminus A_{t+r}} = y_{B_{t+r} \setminus A_{t+r}} | x_{V_t \setminus A_t}, x_{A_{\leq t}}) \end{aligned} \tag{5.47}$$

$$\begin{aligned} =& \delta_{x_{A_{t+r} \cap B_{t+r}}}(y_{A_{t+r} \cap B_{t+r}}) Pr(X_{A_{t+r} \setminus B_{t+r}} = x_{A_{t+r} \setminus B_{t+r}}, \\ & X_{A_{t+r} \cap B_{t+r}} = x_{A_{t+r} \cap B_{t+r}}, \\ & X_{B_{t+r} \setminus A_{t+r}} = y_{B_{t+r} \setminus A_{t+r}} | x_{V_t \setminus A_t}, x_{A_{\leq t}}) \end{aligned} \tag{5.48}$$

$$\begin{aligned} =& \delta_{x_{A_{t+r} \cap B_{t+r}}}(y_{A_{t+r} \cap B_{t+r}}) \\ & p_{A_{t+r} \cup B_{t+r}}(x_{A_{t+r} \setminus B_{t+r}}, x_{A_{t+r} \cap B_{t+r}}, y_{B_{t+r} \setminus A_{t+r}} | x_{V_t \setminus A_t}, x_{A_{\leq t}}) \end{aligned} \tag{5.49}$$

Consequently the five situation above imply the following for the probability of co-occurrence:

1. $A_{t+r} = B_{t+r}$ and $x_{A_{t+r}} = y_{B_{t+r}}$ implies

$$Pr(X_{A_{t+r}} = x_{A_{t+r}}, X_{B_{t+r}} = y_{B_{t+r}} | x_{V_t \setminus A_t}, x_{A_{\leq t}}) = p_{A_{t+r}}(x_{A_{t+r}} | x_{V_t \setminus A_t}, x_{A_{\leq t}}) \tag{5.50}$$

$$= p_{B_{t+r}}(y_{B_{t+r}} | x_{V_t \setminus A_t}, x_{A_{\leq t}}). \tag{5.51}$$

[8] This conditioning can also be removed in the following calculation. However, since we are only interested in probabilities under these conditions in this section we keep it.

2. $A_{t+r} = B_{t+r}$ and $x_{A_{t+r}} \neq y_{B_{t+r}}$ implies:

$$Pr(X_{A_{t+r}} = x_{A_{t+r}}, X_{B_{t+r}} = y_{B_{t+r}} | x_{V_t \setminus A_t}, x_{A_{\leq t}}) = 0 \tag{5.52}$$

3. $A_{t+r} \neq B_{t+r}$, $A_{t+r} \cap B_{t+r} \neq \varnothing$ and $x_{A_{t+r} \cap B_{t+r}} = y_{A_{t+r} \cap B_{t+r}}$ implies

$$Pr(X_{A_{t+r}} = x_{A_{t+r}}, X_{B_{t+r}} = y_{B_{t+r}} | x_{V_t \setminus A_t}, x_{A_{\leq t}}) \tag{5.53}$$

$$= p_{A_{t+r} \cup B_{t+r}}(x_{A_{t+r} \setminus B_{t+r}}, x_{A_{t+r} \cap B_{t+r}}, y_{B_{t+r} \setminus A_{t+r}} | x_{V_t \setminus A_t}, x_{A_{\leq t}}) \tag{5.54}$$

4. $A_{t+r} \neq B_{t+r}$, $A_{t+r} \cap B_{t+r} \neq \varnothing$ and $x_{A_{t+r} \cap B_{t+r}} \neq y_{A_{t+r} \cap B_{t+r}}$ implies

$$Pr(X_{A_{t+r}} = x_{A_{t+r}}, X_{B_{t+r}} = y_{B_{t+r}} | x_{V_t \setminus A_t}, x_{A_{\leq t}}) = 0 \tag{5.55}$$

5. $A_{t+r} \cap B_{t+r} = \varnothing$ implies

$$Pr(X_{A_{t+r}} = x_{A_{t+r}}, X_{B_{t+r}} = y_{B_{t+r}} | x_{V_t \setminus A_t}, x_{A_{\leq t}}) \tag{5.56}$$

$$= p_{A_{t+r}, B_{t+r}}(x_{A_{t+r}}, y_{B_{t+r}} | x_{V_t \setminus A_t}, x_{A_{\leq t}}). \tag{5.57}$$

5.5.4 The problems of exhaustion and mutual-exclusion

In order to define our notion of perception we need to define a suitable conditional probability distribution over the next time-slices or the futures of the co-perception entities $\mathfrak{S}(x_A, t)$. Intuitively, we want to know with what probability which entity occurs at the next time-slice in order to classify the environments accordingly. We are then only interested in cases where one of the co-perception entities' (non-empty) futures actually occurs. Other situations should not be taken into consideration since they do not concern the co-perception entities. However, in general it is possible that none of the co-perception entities occurs. This also poses a formal problem since the probability distribution should range over a set of possible outcomes such that one of them always occurs i.e. the sum over the probabilities of all outcomes must be one. In other words the outcomes must be exhaustive. Another requirement is that the outcomes are mutually exclusive, i.e. only one (and with the previous requirement exactly one) of the outcomes occurs. This property is also not satisfied in general for co-perception entities. In the following we will take a short look at how to construct a probability distribution over a set of events that is not exhaustive. This construction will also require that the events are mutually exclusive. The construction is basically elementary and well known probability theory. We expose it here in some detail to show why we require mutual exclusion and exhaustion and how these requirements are implicit in the perception-action loop.

Consider again the simple case where $\mathfrak{S}(x_A, t) = \{x_A, y_B\}$. From the multivariate Markov chain we know the probabilities for each of them given the identical past and the environment i.e. we know

$$q := p_{A_{t<}}(x_{A_{t<}} | x_{V_t \setminus A_t}, x_{A_{\leq t}}) \tag{5.58}$$

and

$$r := p_{B_{t<}}(y_{B_{t<}} | x_{V_t \setminus A_t}, x_{A_{\leq t}}). \tag{5.59}$$

It is not guaranteed that any of the two STPs will occur. It is generally guaranteed if $A_{t<} = B_{t<}$ and $X_{A_{t<}}$ can only take these two values i.e. $\mathcal{X}_{A_{t<}} = \{x_{A_{t<}}, y_{A_{t<}}\}$ or at least if for all $z_{A_{t<}} \in \mathcal{X}_{A_{t<}}$ and environments $x_{V_t \setminus A_t}$ we have

$$p_{A_{t<}}(z_{A_{t<}} | x_{V_t \setminus A_t}, x_{A_{\leq t}}) = 0. \tag{5.60}$$

In all other cases the random variables with indices in $A_{t<} \cup B_{t<}$ can take values that lead to neither $x_{A_{t<}}$ nor $y_{B_{t<}}$ occurring.[9]

In order to state the property of exhaustiveness formally we first define a notation for the probability that *one or more* elements of a set $\{x^k_{A^k}\}_{k \in I}$ of STPs occurs. We here stop conditioning on $x_{V_t \setminus A_t}, x_{A_{\leq t}}$ here for readability. Conditioning all following probabilities does not interfere with the argument.

Definition 79. *Let $\{X_i\}_{i \in V}$ be a multivariate Markov chain with $V = J \times T$ also let $\mathfrak{C} = \{x^k_{A^k}\}_{k \in I}$ be a set of STPs. We then write*

$$Pr\left(\bigcup \mathfrak{C}\right) := Pr\left(\bigcup_{k \in I} \{X_{A^k} = x^k_{A^k}\}\right) := Pr\left(\bigcup_{k \in I} \mathcal{T}(x^k_{A^k})\right). \tag{5.61}$$

Where $\mathcal{T}(x^k_{A^k})$ is the set of trajectories that $x^k_{A^k}$ occurs in (Definition 19).

Then we define that a set of STPs $\mathfrak{C}$ is exhaustive if the probability that one or more of them occur is one.

Definition 80 (Exhaustiveness of a set of STPs)**.** *Let $\{X_i\}_{i \in V}$ be a multivariate Markov chain with $V = J \times T$ also let $\mathfrak{C} = \{x^k_{A^k}\}_{k \in I}$ be a set of STPs. We say that $\mathfrak{C}$ is exhaustive if*

$$Pr\left(\bigcup \mathfrak{C}\right) = 1. \tag{5.62}$$

Else, we call $\mathfrak{C}$ non-exhaustive.

Remark:

- It is important to note that for any subset $A \subseteq V$ of V the set $\mathcal{X}_A$ is an exhaustive set of STPs. An example we have seen in Section 5.5.1 before is the set $\mathcal{M}_{t+1}$ of possible next values of the agent-process. An example we will encounter later in Section 5.6 is $\mathcal{M}_{t<}$, the set of possible futures starting from time $t+1$ of the agent process in a perception action loop.

Since there is no guarantee that the set $\mathfrak{S}(x_A, t)$ of co-perception entities is exhaustive, the question is how to construct a (conditional) probability distribution over a set of non-exhaustive STPs. The standard approach is to use the definition of conditional probabilities of two events E, F (with $p(F) > 0$) of the form:

$$p(E|F) := \frac{p(E \cap F)}{p(F)}. \tag{5.63}$$

[9] For example if we let $C_{t<} := A_{t<} \cup B_{t<}$ and define $z_{C_{t<}}$ in such a way that it differs from both $x_{A_{t<}}$ and $y_{B_{t<}}$ i.e. there exist nodes $i \in A_{t<}, j \in B_{t<}$ (possibly with $i = j$) such that $z_i \neq x_i$ and $z_j \neq y_j$. Then if $p_{C_{t<}}(z_{C_{t<}} | x_{V_t \setminus A_t}, x_{A_{\leq t}}) > 0$ we have $q + r < 1$.

With this we can define the probability of any single STP $x^k_{A^k} \in \mathfrak{C}$ given that one or more of the STPs in a non-exhaustive set $\mathfrak{C}$ occur. At least as long as one or more of the STPs *can* occur. I.e. if

$$\Pr\left(\bigcup \mathfrak{C}\right) > 0. \tag{5.64}$$

we can replace $F \to \bigcup \mathfrak{C}$ and $E \to \{X_{A_k} = x^k_{A^k}\}$ in Eq. (5.63) to get

$$\begin{aligned}
\Pr\left(X_{A_k} = x^k_{A^k} | \bigcup \mathfrak{C}\right) &= \frac{\Pr\left(\{X_{A_k} = x^k_{A^k}\} \cap \bigcup \mathfrak{C}\right)}{\Pr(\bigcup \mathfrak{C})} && (5.65)\\
&= \frac{\Pr\left(\mathcal{T}(x^k_{A^k}) \cap \bigcup \mathfrak{C}\right)}{\Pr(\bigcup \mathfrak{C})} && (5.66)\\
&= \frac{\Pr\left(\mathcal{T}(x^k_{A^k}) \cap \bigcup_{l \in I} \mathcal{T}(x^l_{A^l})\right)}{\Pr(\bigcup \mathfrak{C})} && (5.67)\\
&= \frac{\Pr\left(\mathcal{T}(x^k_{A^k})\right)}{\Pr(\bigcup \mathfrak{C})} && (5.68)\\
&= \frac{\Pr(X_{A_k} = x^k_{A^k})}{\Pr(\bigcup \mathfrak{C})} && (5.69)\\
&= \frac{p_{A_k}(x^k_{A^k})}{\Pr(\bigcup \mathfrak{C})}. && (5.70)
\end{aligned}$$

While conditioning on $\mathfrak{C}$ guarantees that one of the STPs $x^k_{A^k}$ occurs it still does not necessarily result in a probability distribution since the sum over all STPs in $\mathfrak{C}$ may not be equal to one:

$$\begin{aligned}
\sum_{k \in I} \Pr\left(X_{A_k} = x^k_{A^k} | \bigcup \mathfrak{C}\right) &= \frac{\sum_{k \in I} p_{A_k}(x^k_{A^k})}{\Pr\left(\bigcup \mathfrak{C}\right)} && (5.71)\\
&\geqslant 1. && (5.72)
\end{aligned}$$

However, if the STPs in $\mathfrak{C}$ are mutually exclusive this changes.

Definition 81 (Mutual exclusion)**.** *Let $\{X_i\}_{i \in V}$ be a multivariate Markov chain with $V = J \times T$ also let $\mathfrak{C} = \{x^k_{A^k}\}_{k \in I}$ be a set of STPs. We say that $\mathfrak{C}$ is a set of mutually exclusive STPs if for all $k, l \in I$ we have*

$$\Pr(X_{A_k} = x^k_{A^k}, X_{A_l} = x^l_{A^l}) = 0. \tag{5.73}$$

Remark:

- It is important to note that for any subset $A \subseteq V$ of V the set $\mathcal{X}_A$ is a mutually exclusive set of STPs. Both $\mathcal{M}_{t+1}$ and $\mathcal{M}_{t<}$ of the agent process in a perception action loop are examples of this.

If $\mathfrak{C}$ is mutually exclusive then we get:

$$\Pr\left(\bigcup \mathfrak{C}\right) = \sum_{k \in I} p_{A_k}(x^k_{A^k}) \tag{5.74}$$

and

$$\sum_{k\in I} \Pr\left(X_{A_k} = x^k_{A^k} | \bigcup \mathfrak{C}\right) = 1 \tag{5.75}$$

in general. We can then also write:

$$\Pr\left(X_{A_k} = x^k_{A^k} | \bigcup \mathfrak{C}\right) = \frac{p_{A_k}(x^k_{A^k})}{\sum_{l\in I} p_{A_l}(x^l_{A^l})} \tag{5.76}$$

in our usual notation. Then the above defines a probability for each element of the set $\mathfrak{C}$ or equivalently for each index $k \in I$. We can then define the probability distribution $p_{\mathfrak{C}} : \mathfrak{C} \to [0,1]$ via

$$p_{\mathfrak{C}}(x^k_{A^k}) := \frac{p_{A_k}(x^k_{A^k})}{\sum_{l\in I} p_{A_l}(x^l_{A^l})} \tag{5.77}$$

In summary, if we have a set $\mathfrak{C}$ of non-exhaustive but mutually exclusive STPs we now know how to define a probability distribution over them. For reference we put this in a theorem.[10]

Theorem 35 (Probability distribution construction for non-exhaustive but mutually exclusive sets of STPs)**.** *Let $\{X_i\}_{i\in V}$ be a multivariate Markov chain with $V = J \times T$ also let $\mathfrak{C} = \{x^k_{A^k}\}_{k\in I}$ be a set of STPs. If $\mathfrak{C}$ is mutually exclusive (not necessarily exhaustive) and*

$$\sum_{l\in I} p_{A_l}(x^l_{A^l}) > 0 \tag{5.78}$$

we can define a probability distribution $p_{\mathfrak{C}} : \mathfrak{C} \to [0,1]$ via

$$p_{\mathfrak{C}}(x^k_{A^k}) := \frac{p_{A_k}(x^k_{A^k})}{\sum_{l\in I} p_{A_l}(x^l_{A^l})}. \tag{5.79}$$

For convenience we also write this as a probability distribution over an index set of $\mathfrak{S}$:

$$p_{\mathfrak{C}}(k) := p_{\mathfrak{C}}(x^k_{A^k}). \tag{5.80}$$

Proof. Along the lines of the preceding argument. What is missing is the calculation that

$$\Pr\left(\bigcup \mathfrak{C}\right) = \sum_{k\in I} p_{A_k}(x^k_{A^k}) \tag{5.81}$$

follows from mutual exclusion. This is straightforward but tedious. The idea is that since the probability of all intersections of the sets of trajectories $\mathcal{T}(x^k_{A^k})$ vanishes (due to mutual exclusion) we get the same result *as if* all $\mathcal{T}(x^k_{A^k})$ were disjoint. The probability of a union of disjoint sets is the sum over the probabilities of the sets. □

[10] This is not an original theorem of this thesis. We presented the preceding arguments since they help to understand subsequent notions.

For the previous example of a simple co-perception set $\mathfrak{S}(x_A, t) = \{x_A, y_B\}$ this means that if they are mutually exclusive i.e.

$$\Pr(X_{A_{t<}} = x_{A_{t<}}, X_{B_{t<}} = y_{B_{t<}} | x_{V_t \setminus A_t}, x_{A_{\leq t}}) = 0 \tag{5.82}$$

and at least one of their conditional probabilities is positive i.e.

$$q + r > 0 \tag{5.83}$$

we can write $x^1_{A^1} := x_A$ and $x^2_{A^2} = y_B$, and $b \in \{1, 2\}$ to get a conditional probability distribution[11]

$$p_{\mathfrak{S}(x_A,t)}(b | x_{V_t \setminus A_t}, x_{A_{\leq t}}) := \frac{p_{A^b}(x^b_{A^b_{t<}} | x_{V_t \setminus A_t}, x_{A_{\leq t}})}{\sum_{c \in \{1,2\}} p_{A^c}(x^c_{A^c_{t<}} | x_{V_t \setminus A_t}, x_{A_{\leq t}})} \tag{5.84}$$

$$= \frac{p_{A^b}(x^b_{A^b_{t<}} | x_{V_t \setminus A_t}, x_{A_{\leq t}})}{q + r}. \tag{5.85}$$

With such a conditional probability distribution we can define perception in basically the same way as for the perception-action loop.

5.5.5 Co-perception environments

Equation (5.84) is already a step towards entity perception since it is a conditional probability distribution over futures of (co-perception) entities i.e. over things already quite similar to "next time-slices of the entity". There are two remaining problems however. The construction of the conditional probability distribution relies on Theorem 35. For this to apply we need mutual exclusion of the STPs (in this case the co-perception entities) and we need Eq. (5.78) to hold. In general the set of co-perception entities $\mathfrak{S}(x_A, t)$ is *not* mutually exclusive. This will be discussed further in Section 5.5.6. In this section we discuss the second problem. Our solution may seem like it includes a strong requirement on the environments. We then show that this requirement is implicit in the perception-action loop as well.

The second problem is the condition of Eq. (5.78) which in the case of co-perception entities concerns the sum over the probabilities of the next time-slices of the co-perception entities $\mathfrak{S}(x_A, t)$ given the identical past:

$$\sum_{k \in I} p_{A^k}(x^k_{A^k_{t<}} | x_{V_t \setminus A_t}, x_{A_{\leq t}}). \tag{5.86}$$

This is not necessarily greater than zero for all environments $x_{V_t \setminus A_t} \in \mathcal{X}_{V_t \setminus A_t}$. In fact if

$$p_{V_t, A_{\leq t}}(x_{V_t \setminus A_t}, x_{A_{\leq t}}) = 0 \tag{5.87}$$

then

$$p_{A^k}(x^k_{A^k_{t<}} | x_{V_t \setminus A_t}, x_{A_{\leq t}}) \tag{5.88}$$

[11] Since all probabilities involved are conditioned on the same STPs the above argument and Theorem 35 hold equally for conditional probabilities.

is not even defined. To be able to use Theorem 35 we must therefore require of environments $x_{V_t \setminus A_t}$ that can be classified that

$$p_{V_t, A_{\leq t}}(x_{V_t \setminus A_t}, x_{A_{\leq t}}) > 0 \tag{5.89}$$

and that there is at least one element $x^k_{A^k_{t<}} \in \mathfrak{S}(x_A, t)$ with

$$p_{A^k_{t<}, V_t, A_{\leq t}}(x^k_{A^k_{t<}}, x_{V_t \setminus A_t}, x_{A_{\leq t}}) > 0 \tag{5.90}$$

We can summarise these two conditions as the condition that there exists $x^k_{A^k} \in \mathfrak{S}(x_A, t)$ with

$$p_{A^k, V_t}(x^k_{A^k}, x_{V_t \setminus A_t}) > 0 \tag{5.91}$$

where we used that $A^k_{\leq t} = A_{\leq t}$. We can also get rid of the need for the existence quantifier by writing this condition as:

$$\Pr\left(\bigcup \mathfrak{S}(x_A, t) \cap \{X_{V_t \setminus A_t} = x_{V_t \setminus A_t}\}\right) > 0. \tag{5.92}$$

We call the subset of such environments the *co-perception environments*.

Definition 82 (Co-perception environments)**.** *Let $\{X_i\}_{i \in V}$ be a multivariate Markov chain with $V = J \times T$ and entity set $\mathfrak{E}$. Let $x_A \in \mathfrak{E}$ be an entity with non-empty time-slices at t and $t+1$ and $\mathfrak{S}(x_A, t)$ its co-perception entities. Then define the associated co-perception environments $\mathcal{X}^{\mathfrak{S}}_{V_t \setminus A_t} \subseteq \mathcal{X}_{V_t \setminus A_t}$ by*

$$\mathcal{X}^{\mathfrak{S}}_{V_t \setminus A_t} := \{\bar{x}_{V_t \setminus A_t} \in \mathcal{X}_{V_t \setminus A_t} : \exists y_B \in \mathfrak{S}(x_A, t), p_{B, V_t \setminus A_t}(y_B, \bar{x}_{V_t \setminus A_t}) > 0\}. \tag{5.93}$$

Remark:

- The co-perception environments of a co-perception set $\mathfrak{S}(x_A, t)$ are then the spatial patterns $\mathcal{X}_{V_t \setminus A_t}$ at t that can co-occur with at least one co-perception environment.

It may seem like a (too) strong requirement that the co-perception environments are compatible with an entire entity including its (entire) future. This also seems strange from a causal perspective. One way to interpret this is to say that the co-perception environments are just the environments that are ever going to be classified by an entity. Whenever an environment that is not in $\mathcal{X}^{\mathfrak{S}}_{V_t \setminus A_t}$ occurs at t together with the identical past $x_{A_{\leq t}}$ there will not be any perception since no entity with the identical past will be there at $t+1$. Whenever there is an entity with the identical past at $t+1$ the environment is in $\mathcal{X}^{\mathfrak{S}}_{V_t \setminus A_t}$.

At the same time this assumption is also implicit in the perception-action loop. Say (as is the case in the perception-action loop) the futures of the co-perception entities are just the possible values of a set C of random variables i.e. if there exists $C \subseteq V_{t<}$ such that

$$\mathfrak{S}(x_A, t)_{t<} := \{y_{B_{t<}} : y_B \in \mathfrak{S}(x_A, t)\} = \mathcal{X}_C. \tag{5.94}$$

Then futures of the co-perception entities are exhaustive and mutually exclusive. In that case it turns out that it is sufficient to require that the co-perception environments can co-occur with the identical past $x_{A_{\leq t}}$. The requirement that it can co-occur with at least one future of a co-perception entity is then automatically satisfied and vice versa. So if $\mathfrak{S}(x_A, t)_{t<} = \mathcal{X}_C$ it is sufficient to define

$$\mathcal{X}^{\mathfrak{S}}_{V_t \setminus A_t} := \{\bar{x}_{V_t \setminus A_t} \in \mathcal{X}_{V_t \setminus A_t} : p_{A_{\leq t}, V_t \setminus A_t}(x_{A_{\leq t}}, \bar{x}_{V_t \setminus A_t}) > 0\}. \tag{5.95}$$

This condition *always* needs to be satisfied for probabilities $p(.|x_{A_{\leq t}}, \bar{x}_{V_t \setminus A_t})$, that condition on the identical past $x_{A_{\leq t}}$ and environment $\bar{x}_{V_t \setminus A_t}$ to be defined. To see that the two sets of Eqs. (5.93) and (5.95) are equal recall with Eq. (5.92) that if $\mathfrak{S}(x_A, t)_{t<} = \mathcal{X}_C$ an environment $\bar{x}_{V_t \setminus A_t}$ is in $\mathcal{X}^{\mathfrak{S}}_{V_t \setminus A_t}$ if

$$\Pr\left(\bigcup \mathcal{X}_C \cap \{X_{A_{\leq t}} = x_{A_{\leq t}}\} \cap \{X_{V_t \setminus A_t} = \bar{x}_{V_t \setminus A_t}\}\right) > 0. \tag{5.96}$$

Using the mutual exclusion and exhaustiveness of the random variable X_C we note that

$$\begin{aligned}
&\Pr\left(\bigcup \mathcal{X}_C \cap \{X_{A_{\leq t}} = x_{A_{\leq t}}\} \cap \{X_{V_t \setminus A_t} = \bar{x}_{V_t \setminus A_t}\}\right) && (5.97)\\
=&\sum_{\hat{x}_C \in \mathcal{X}_C} \Pr\left(\{X_C = \hat{x}_C\} \cap \{X_{A_{\leq t}} = x_{A_{\leq t}}\} \cap \{X_{V_t \setminus A_t} = \bar{x}_{V_t \setminus A_t}\}\right) && (5.98)\\
=&\sum_{\hat{x}_C \in \mathcal{X}_C} p_{C, A_{\leq t}, V_t \setminus A_t}(\hat{x}_C, x_{A_{\leq t}}, \bar{x}_{V_t \setminus A_t}) && (5.99)\\
=&p_{A_{\leq t}, V_t \setminus A_t}(x_{A_{\leq t}}, \bar{x}_{V_t \setminus A_t}). && (5.100)
\end{aligned}$$

Where we used mutual exclusion from the second to the third line and exhaustiveness from the fourth to the fifth. So the two sets of Eqs. (5.93) and (5.95) are identical in this case. Without mutual exclusion and exhaustiveness the two sets are not equal.

More intuitively this can be understood by noting that assuming that the entities exhaust a set of future random variables X_C means that there is a future of a co-perception entity in *every* trajectory compatible with $x_{A_{\leq t}}$. Since all environments that are compatible with $x_{A_{\leq t}}$ must occur in one of those trajectories Eq. (5.93) is automatically satisfied. Our definition of co-perception environments therefore does not contain an additional assumption compared to cases where futures of entities exhaust sets of random variables like in the case of the perception-action loop or the models of biological individuals of Krakauer et al. (2014).

5.5.6 Non-interpenetration and mutual exclusion

In this section we define the formal assumptions of general non-interpenetration and (past specific) non-interpenetration for entity sets. This leads to mutual-exclusion of entities in ways that enable a unique definition of our notion of entity perception. Without these assumptions we can still define entity perception but there will an arbitrary choice involved which influences the perceptions (we will see this in Section 5.5.8).

So different choices mean that the extracted perceptions are different. This is not a desirable situation since we are making this choice. The goal of this thesis is, however, that all the notions only depend on the multivariate Markov chain itself. Non-interpenetration is therefor a desirable property.

General non-interpenetration requires that any two STPs x_A, y_B that partly occupy the same random variables i.e.

$$A \cap B \neq \varnothing \tag{5.101}$$

never co-occur i.e.

$$\Pr(X_A = x_A, X_B = y_B) = 0. \tag{5.102}$$

This notion of non-interpenetration treats spatial and temporal overlap (we call $A \cap B$ the *overlap*) equally. There is a philosophical debate about whether interpenetration is possible for real objects (see e.g. Gilmore (2014)). Here we only want to suggest that non-interpenetration may be a reasonable assumption for entities. Non-interpenetration says that the same spatiotemporal region cannot be occupied by two different entities. This is intuitively true for solid objects. It is however somewhat problematic when we think of multicellular organisms as entities that may contain cells that are also entities by themselves. In that case these cells would be parts of two different entities and violate non-interpenetration. It is also possible that the cells by themselves are actually not entities according to some other entity criterion. We also note that on the level of cells we do not have interpenetration. Two different cells never occupy the same spatiotemporal region. A cell may divide, but that is one spatiotemporal entity dividing *spatially* into two at some point in time. This suggests that there may be some levels of organisation or hierarchies[12] of entities involved. We will not further discuss this here. Instead we only note that the formal property of general non-interpenetration relates to our notion of perception by providing mutual exclusion of co-perception entities. We now state the definition of general non-interpenetration for future reference.

Definition 83 (General non-interpenetration)**.** *Let $\{X_i\}_{i \in V}$ be a multivariate Markov chain with $V = J \times T$. An entity set $\mathfrak{E} \subseteq \bigcup_{B \subseteq V} \mathcal{X}_B$ satisfies general non-interpenetration or is generally non-interpenetrating if for all $y_B, z_C \in \mathfrak{E}$ with $y_B \neq z_C$ we have*

$$B \cap C \neq \varnothing \Rightarrow \Pr(X_B = y_B, X_C = z_C) = 0. \tag{5.103}$$

It turns out that general non-interpenetration is not necessary for our purposes. We require only that all co-perception entities are mutually exclusive. For this it is already sufficient that for two entities x_A, y_B if there is a time $t \in T$ such that they have identical *pasts* up to t

$$x_{A_{\leq t}} = y_{B_{\leq t}} \tag{5.104}$$

but are then different at some time in the future

$$x_{A_{t<}} \neq y_{B_{t<}} \tag{5.105}$$

[12] Whether the disintegration hierarchies are related is beyond the scope of this thesis but an interesting avenue to pursue in the future.

must be mutually exclusive *given that their pasts occurred*:

$$\Pr(X_{A_{t<}} = x_{A_{t<}}, X_{B_{t<}} = y_{B_{t<}} | x_{A_{\leq t}}) = 0. \tag{5.106}$$

This means that there cannot be two different entities which are identical up to some point in time t and then, in the same *single* trajectory (with positive probability), at some point "reveal" their difference. If entities with identical pasts *ever* reveal their difference they must be in different trajectories i.e. they must be mutually exclusive. We could call this "past specific non-interpenetration" but since we only need this notion outside of this section we will just refer to it as non-interpenetration.

Definition 84 (Non-interpenetration)**.** *Let $\{X_i\}_{i\in V}$ be a multivariate Markov chain with $V = J \times T$. An entity set $\mathfrak{E} \subseteq \bigcup_{B\subseteq V} \mathcal{X}_B$ satisfies non-interpenetration or is non-interpenetrating if for all $y_B, z_C \in \mathfrak{E}$ we have*

$$\begin{aligned} \exists t \in T : y_{B_{\leq t}} = z_{C_{\leq t}} \text{ and } y_{B_{t<}} \neq z_{C_{t<}} \\ \Rightarrow \Pr(X_{B_{t<}} = y_{B_{t<}}, X_{B_{t<}} = z_{C_{t<}} | y_{B_{\leq t}}) = 0. \end{aligned} \tag{5.107}$$

Remark:

- We note here that non-interpenetration is not necessarily satisfied by ι-entities as we will see in Section 6.6.

Non-interpenetration implies that co-perception entities are mutually exclusive:

Theorem 36. *Let $\{X_i\}_{i\in V}$ be a multivariate Markov chain with $V = J \times T$ and entity set $\mathfrak{E}$. Let $x_A \in \mathfrak{E}$ be an entity with non-empty time-slices at t and $t+1$ and $\mathfrak{S}(x_A, t)$ its co-perception entities. If $\mathfrak{E}$ satisfies non-interpenetration then $\mathfrak{S}(x_A, t)$ is mutually exclusive.*

Proof. Let $y_B, z_C \in \mathfrak{S}(x_A, t)$ with $y_B \neq z_C$. Then they have identical pasts and so we have $y_{B_{\leq t}} = z_{C_{\leq t}}$. From non-interpenetration we then get

$$\Pr(X_B = y_B, X_C = z_C) = 0. \tag{5.108}$$

□

Remark:

- Note that non-interpenetration does not imply anything about exhaustiveness. We can have non-interpenetrating co-perception entities that are not exhaustive. But since we know how to define a conditional probability distribution for non-exhaustive sets of STPs (Theorem 35) this is not a problem.

This means that under non-interpenetration we can always define a conditional probability distribution over the entire futures of a set $\mathfrak{S}(x_A, t)$ of co-perception entities.

Definition 85 (Co-perception morph)**.** *Let $\{X_i\}_{i\in V}$ be a multivariate Markov chain with $V = J \times T$ and a non-interpenetrating entity set $\mathfrak{E}$. Let $x_A \in \mathfrak{E}$ be an entity*

with non-empty time-slices at t and $t+1$ and $\mathfrak{S}(x_A, t) = \{x^k_{A^k}\}_{k\in I}$ its co-perception entities. Furthermore let $\mathcal{X}^{\mathfrak{S}}_{V_t\setminus A_t} \subseteq \mathcal{X}^{\mathfrak{S}}_{V_t\setminus A_t}$ be the set of co-perception environments. Then the conditional probability distribution $p_{\mathfrak{S}} : \mathfrak{S} \to [0,1]$

$$p_{\mathfrak{S}(x_A,t)}(k|x_{V_t\setminus A_t}, x_{A_{\leq t}}) := \frac{p_{A^k}(x^k_{A^k_{t<}}|x_{V_t\setminus A_t}, x_{A_{\leq t}})}{\sum_{l\in I} p_{A^l}(x^l_{A^l_{t<}}|x_{V_t\setminus A_t}, x_{A_{\leq t}})} \tag{5.109}$$

is well defined and we call it the co-perception morph.

The co-perception morph is a conditional probability distribution over the entire futures of the co-perception entities. It can be used to partition the co-perception environments by assigning environments to the same block / perception if the lead to the same co-perception morph. However, we wanted to define perception that occurs from t to $t+1$. In the co-perception morph there may be entities that only start differing far in the future. If two environments have different influences only on these two entities and equal influence on all other co-perception entities the co-perception morph still distinguishes the two environments at t already. This does not seem like a good definition of perception from t to $t+1$. We therefore partition all co-perception entities according to their next time-slices in the next section.

5.5.7 Branching partition

In this section we will present a partition of the co-perception entities called the branching partition. This will put all entities into the same blocks (called *branches*) that have identical time-slices at $t+1$. This construction is intuitive as it ignores differences between co-perception entities that only become apparent at times later than $t+1$. Since we are interested in the perceptions that happen in the transition from t to $t+1$ such differences should be ignored. The branches (blocks) of the branching partition are therefore an even better analogue to the "possible next time-slices" than the set of co-perception entities directly.

First some more detail: at each transition from time-step t to $t+1$ the co-perception entities $\mathfrak{S}(x_A, t)$ split up into sets of entities that are identical up to $t+1$ (we will call these sets the *branches*). Only one of these sets is the set $\mathfrak{S}(x_A, t+1)$. For example an entity $y_B \in \mathfrak{S}(x_A, t)$ with the same past up to t but with a different time-slice at $t+1$ i.e. $y_{B_{t+1}} \neq x_{A_{t+1}}$ is part of a different branch. In that case this branch is $\mathfrak{S}(y_B, t+1)$ and we have $\mathfrak{S}(y_B, t+1) \cap \mathfrak{S}(x_A, t+1) = \varnothing$. In summary then the dynamics of the system split up the co-perception entities of x_A up to t into disjoint sets (the branches) of entities with identical pasts up to $t+1$. We can then interpret the branches at the time $t+1$ as the distinctions among the co-perception entities that are revealed at time $t+1$. Further distinctions among the co-perception entities are only revealed at later times. This also means that these are *all* differences that could possibly be due to the influence of the environment at t and that show their effect at $t+1$ (not later). In this way the perceptions at t should also be defined with respect to these branches. We call the partition that is defined via the identification of entities in $\mathfrak{S}(x_A, t)$ that are identical up to $t+1$ the *branching partition*.

Definition 86 (Branching partition)**.** *Let $\{X_i\}_{i\in V}$ be a multivariate Markov chain with $V = J \times T$ and entity set $\mathfrak{E}$. Let $x_A \in \mathfrak{E}$ be an entity with non-empty time-slices at t and $t+1$ and $\mathfrak{S}(x_A, t)$ its co-perception entities. Then define the branching partition $\eta(x_A, t)$ of $\mathfrak{S}(x_A, t)$ as the partition induced by the equivalence classes of the equivalence relation*

$$\begin{aligned} y_B \sim z_C \\ \Leftrightarrow y_{B_{t+1}} = z_{C_{t+1}}, \end{aligned} \tag{5.110}$$

where $y_B, z_C \in \mathfrak{S}(x_A, t)$.

Remark:

- The definition of the branching partition can easily be generalised to more than one time-step into the future. Instead of requiring equality at $t+1$ we can require equality for the next r time-steps:

$$\begin{aligned} y_B \sim z_C \\ \Leftrightarrow y_{B_{t+1:t+r}} = z_{C_{t+1:t+r}} \end{aligned} \tag{5.111}$$

  This leads to a partition of $\mathfrak{S}(x_A, t)$ which is a refinement of $\eta(x_A, t)$. The branches of $t+1$ are further partitioned according to the equality of the contained entities at $t+2$, and similarly at each subsequent time-step. This may be used to construct a kind of multi-time-step perception which is more precise than one-time-step perception. Here we focus only on the one-step-perception, nonetheless all further notions are easily adapted to the multi-step case.

The branches of the branching partition are the final analogue notion of "the entity's next possible time-slices". We then define a conditional probability distribution over the branches (called the branch-morph) and classify the co-perception environments accordingly.

5.5.8 Branch-morph

Given the branching partition $\eta(x_A, t)$ for a non-interpenetrating entity set we can then define a conditional probability distribution over the branches by just summing up the probabilities of all entities in each branch (remember that they are all mutually exclusive) to get the probability of a branch i.e. write for every block $b \in \eta(x_A, t)$:

$$p(b|\hat{x}_{V_t\setminus A_t}, x_{A_{\leq t}}) := \sum_{y_B \in b} p_{B_{t<}, V_t\setminus A_t}(y_{B_{t<}}|\hat{x}_{V_t\setminus A_t}, x_{A_{\leq t}}). \tag{5.112}$$

Since the branches are also mutually exclusive (because all co-perception entities are mutually exclusive) we can divide by the sum of probabilities of the branches to get a probability distribution.

$$p_{\eta(x_A,t)}(b|\hat{x}_{V_t\setminus A_t}, x_{A_{\leq t}}) := \frac{p(b|\hat{x}_{V_t\setminus A_t}, x_{A_{\leq t}})}{\sum_{c\in\eta(x_A,t)} p(c|\hat{x}_{V_t\setminus A_t}, x_{A_{\leq t}})}. \tag{5.113}$$

This is the idea behind Definition 87 below.

However, we do not necessarily need the mutual exclusion of the *entire* set of co-perception entities if we want to define *some* perception. For this we can use a subset $\zeta(x_A, t) \subset \mathfrak{S}(x_A, t)$ with $x_A \in \zeta(x_A, t)$ of mutually exclusive co-perception entities. We can then still use the branching partition on this subset and define perception in the same way as below by replacing $\mathfrak{S}(x_A, t)$ with $\zeta(x_A, t)$ everywhere.

However, for a set of co-perception entities that is not naturally mutually exclusive as in the case of non-interpenetration there are many possible choices of such subsets which lead to different perceptions. For example say $\mathfrak{S}(x_A, t) = \{x_A, y_B, z_C\}$. Then if they are not all mutually exclusive, we can have that x_A and y_B are mutually exclusive and x_A and z_C are mutually exclusive but y_B and z_C are *not* mutually exclusive. So then to get a proxy $\zeta(x_A, t)$ of $\mathfrak{S}(x_A, t)$ (consisting only of mutually exclusive co-perception entities) we can either choose $\zeta(x_A, t) = \{x_A, y_B\}$ or $\zeta(x_A, t) = \{x_A, z_C\}$. These choices will in general lead to different perceptions such that perception is not unique if we don't have mutual exclusion of all co-perception entities. So non-interpenetration is an attractive property for entity sets if we want to use our notion of perception since it allows us to uniquely define it.

We now finally state the definition of the branch-morph, the co-perception environment partition and the perceptions which are just the blocks of the co-perception environment partition.

Definition 87 (Branch-morph)**.** *Let $\{X_i\}_{i \in V}$ be a multivariate Markov chain with index set $V = J \times T$ and entity set $\mathfrak{E}$. Let $x_A \in \mathfrak{E}$ be an entity with non-empty time-slices at t and $t+1$ and $\mathfrak{S}(x_A, t)$ its co-perception entities and $\eta(x_A, t)$ the branching partition. Furthermore, let $\mathcal{X}^{\mathfrak{S}}_{V_t \setminus A_t} \subseteq \mathcal{X}_{V_t \setminus A_t}$ be the associated co-perception environments. Also write for every block $b \in \eta(x_A, t)$:*

$$p(b|\hat{x}_{V_t \setminus A_t}, x_{A_{\leq t}}) := \sum_{y_B \in b} p_{B_{t<}, V_t \setminus A_t}(y_{B_{t<}}|\hat{x}_{V_t \setminus A_t}, x_{A_{\leq t}}). \tag{5.114}$$

Then for each $\hat{x}_{V_t \setminus A_t} \in \mathcal{X}^{\mathfrak{S}}_{V_t \setminus A_t}$ we define the branch-morph over $\eta(x_A, t)$ as the probability distribution $p_{\eta(x_A,t)}(.|\hat{x}_{V_t \setminus A_t}, x_{A_{\leq t}}) : \eta(x_A, t) \to [0, 1]$ with

$$p_{\eta(x_A,t)}(b|\hat{x}_{V_t \setminus A_t}, x_{A_{\leq t}}) := \frac{p(b|\hat{x}_{V_t \setminus A_t}, x_{A_{\leq t}})}{\sum_{c \in \eta(x_A,t)} p(c|\hat{x}_{V_t \setminus A_t}, x_{A_{\leq t}})}, \tag{5.115}$$

for all $b \in \eta(x_A, t)$.

With the branch-morph we can then define, as expected, the perceptions as equivalence classes of the co-perception environments with respect to the associated branch-morph. First we define a partition of the co-perception *environments* called the co-perception environment partition. The perceptions are then the blocks of this partition.

Definition 88. *Let $\{X_i\}_{i \in V}$ be a multivariate Markov chain with index set $V = J \times T$ and entity set $\mathfrak{E}$. Let $x_A \in \mathfrak{E}$ be an entity with non-empty time-slices at t and $t+1$ and $\mathfrak{S}(x_A, t)$ its co-perception entities and $\eta(x_A, t)$ the branching partition. Furthermore,*

let $\mathcal{X}^{\mathfrak{S}}_{V_t \setminus A_t} \subseteq \mathcal{X}_{V_t \setminus A_t}$ *be the associated co-perception environments. Then define the co-perception environment partition* $\pi^{\mathfrak{S}}(x_A, t)$ *of* $\mathcal{X}^{\mathfrak{S}}_{V_t \setminus A_t}$ *as the partition induced by the equivalence classes of the equivalence relation*

$$\begin{aligned} &\hat{x}_{V_t \setminus A_t} \sim \bar{x}_{V_t \setminus A_t} \\ &\qquad \Leftrightarrow \forall b \in \eta(x_A, t) : p_{\eta(x_A,t)}(b | \hat{x}_{V_t \setminus A_t}, x_{A_{\leq t}}) = p_{\eta(x_A,t)}(b | \bar{x}_{V_t \setminus A_t}, x_{A_{\leq t}}). \end{aligned} \tag{5.116}$$

Remarks:

- This means all associated co-perception environments in the same block of $\pi^{\mathfrak{S}}(x_A, t)$ have the same branch-morph. In other words they lead to the same branch of entity futures (i.e. the same future branch) with the same probabilities. Then all elements of these environment blocks have identical effects on the future branches and these branches cannot distinguish between environments within the blocks.

Definition 89 (Perceptions)**.** *Let* $\{X_i\}_{i \in V}$ *be a multivariate Markov chain with index set* $V = J \times T$ *and entity set* $\mathfrak{E}$*. Let* $x_A \in \mathfrak{E}$ *be an entity with non-empty time-slices at* t *and* $t+1$ *and* $\mathfrak{S}(x_A, t)$ *its co-perception entities. Furthermore, let* $\mathcal{X}^{\mathfrak{S}}_{V_t \setminus A_t} \subseteq \mathcal{X}_{V_t \setminus A_t}$ *be the associated co-perception environments and* $\pi^{\mathfrak{S}}(x_A, t)$ *its co-perception environment partition.*

Then the blocks of $\pi^{\mathfrak{S}}(x_A, t)$ *are called the perceptions of* x_A *at* t*.*

5.6 Entity action and perception in the perception-action loop

We now show that agent-environment systems as modelled by the perception-action loop are multivariate Markov chains containing a specific choice of entity sets.

In this section we interpret the perception-action loop from the perspective of our own concept of agents. We identify an entity set and the subset of the entity set that are agents from the perception-action loop perspective. Both of these choices turn out to be nonrestrictive. The perception-action loop itself puts no strong constraints on the "agents" (often only referred to as *systems*) it models. Its Bayesian network structure of Fig. 5.4 is also compatible with two independent and identically distributed random variables. Accordingly there have been attempts to identify further restrictions or measures that quantify further distinctions among such systems. The most relevant in our context is the measure of *autonomy* for an agent in the perception-action loop (Bertschinger et al., 2008). We show how our notion of actions is sufficient for the property of non-heteronomy which is part of the requirement for autonomy. We also show that, as expected due to our construction, our notion of perception specialises to the notion of perception in the perception action (Section 3.3.6). The main point of this chapter is to show how our notion of an acting and perceiving entity can be seen as a generalisation of (partly

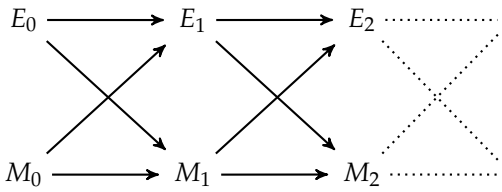


Figure 5.4: First time-steps of the Bayesian network of the perception-action loop. The processes represent environment $\{E_t\}_{t\in T}$, and agent memory $\{M_t\}_{t\in T}$.

autonomous) agents as they are modelled in perception-action loops. Our notions are more general because they are well defined for non-exhaustive and extent varying entity-sets.

Recall that an agent in a perception-action loop (Definition 43) is a stochastic process $\{M_t\}_{t\in T}$ interacting with an environment process $\{E_t\}_{t\in T}$. If we want to make the interactions explicit we can use the extended perception-action loop of Definition 46. In this case we also have an action process $\{A_t\}_{t\in T}$ and a sensor process $\{S_t\}_{t\in T}$. For convenience we again show the Bayesian network (of the non-extended version) in Fig. 5.4.

In the perception-action loop each trajectory x_V is considered to consist of a time-evolution m_T of the agent and a time-evolution of the environment e_T. The agent therefore occurs in every trajectory and occupies the same degree of freedom in every trajectory. According to our working definition (Definition 64) all agents are entities and according to our concept of entities (Definition 65), entities are STPs. Each of the time-evolutions m_T is a STP in the perception-action loop. We can then define the entity set $\mathfrak{E}$ of a perception-action loop as the set of time-evolutions of the agent process i.e.

$$\mathfrak{E}^{PA} := \{m_T \in \prod_{t\in T} \mathcal{M}_t\}. \tag{5.117}$$

With this definition every time-evolution of the agent process corresponds to an entity and every such entity corresponds to the time-evolution of an agent (or *the* agent). Similarly, we can define entities for the environments and add them to $\mathfrak{E}^{PA}$. The symmetry of the perception-action loop makes no a priori difference between agents and environments. We will focus on the agent process here and do not need environment entities.

Compared to ι-entities (Definition 74) the entity set $\mathfrak{E}^{PA}$ is not very restrictive. Even if each M_t for $t \in T$ is a independently distributed random variable the sequences m_T would still be considered time evolutions of entities. Furthermore, each of these entities would be considered an agent in this picture. In order to introduce stronger conditions efforts have been made to distinguish informationally closed (Bertschinger et al., 2006) and autonomous (Bertschinger et al., 2008) agents. Our own concept of agents also puts stronger constraints on the notion of an entities. We require actions, perception, and goal-directedness from entities that can be counted as agents. We have not defined a notion of goal-directedness, but the notions of entity

actions (Section 5.4) and entity perceptions (Section 5.5) can be used for the perception-action loop entity set $\mathfrak{E}^{PA}$. We will consider this next and see that our more general requirements of actions corresponds to a requirement by Bertschinger et al. in the case of perception-action loop entity sets.

5.6.1 Entity actions in the perception-action loop

According to our definition (Definition 76) an entity x_A performs an action at time t in a trajectory x_V if there is a co-action entity y_B occurring in a co-action trajectory y_V with $x_{V_t \setminus A_t} = y_{V_t \setminus B_t}$. In the case of the perception action loop we can write every trajectory as a pair (m_T, e_T) where m_T is an entity. The entity m_T then performs an action at time t in trajectory (m_T, e_T) with $p_{M_T,E_T}(m_T, e_T) > 0$ if there is an entity $\bar{m}_t$ such that

- $\bar{m}_T$ occurs in $(\bar{m}_T, \bar{e}_T) \neq (m_T, e_T)$ with $p_{M_T,E_T}(\bar{m}_T, \bar{e}_T) > 0$,
- at t the entities m_T and $\bar{m}_T$ occupy the same random variables, which is the case for all entities in $\mathfrak{E}^{PA}$,
- at t the environments of m_T and $\bar{m}_T$ are identical: $e_t = \bar{e}_t$,
- at $t+1$ the entities are different: $m_{t+1} \neq \bar{m}_{t+1}$.

Since all entities occupy the same random variables we can only have value actions in the perception-action loop.

We now show that these conditions can be related to conditions for autonomous systems/agents proposed by Bertschinger et al. (2008). We can say that the more entities perform actions at t the higher is potentially the non-heteronomy component of a measure of autonomy proposed by Bertschinger et al..

If we assume that these conditions are fulfilled at some time t for two entities $m_T, \bar{m}_T$ we can derive that the conditional entropy $\mathrm{H}(M_{t+1}|E_t)$ of the next agent state given the current environment state is greater than zero:

$$\mathrm{H}(M_{t+1}|E_t) > 0. \tag{5.118}$$

Proof. From $p_{M_T,E_T}(m_T, e_T) > 0$ and $p_{M_T,E_T}(\bar{m}_T, \bar{e}_T) > 0$ it directly follows that $p_{M_{t+1}}(m_{t+1}|e_t) > 0$, $p_{M_{t+1}}(\bar{m}_{t+1}|e_t) > 0$ and $p_{E_t}(e_t) > 0$. Then:

$$\mathrm{H}(M_{t+1}|E_t) := -\sum_{\hat{e}_t \in \mathcal{E}_t} p_{E_t}(\hat{e}_t) \sum_{\hat{m}_{t+1} \in \mathcal{M}_{t+1}} p_{M_{t+1}}(\hat{m}_{t+1}|\hat{e}_t) \log p_{M_{t+1}}(\hat{m}_{t+1}|\hat{e}_t) \tag{5.119}$$

$$\geqslant -p_{E_t}(e_t) \Big(p_{M_{t+1}}(m_{t+1}|e_t) \log p_{M_{t+1}}(m_{t+1}|e_t) + p_{M_{t+1}}(\bar{m}_{t+1}|e_t) \log p_{M_{t+1}}(\bar{m}_{t+1}|e_t) \Big) \tag{5.120}$$

$$> 0. \tag{5.121}$$

□

We can also see from this that the more entities perform actions at t the more terms in Eq. (5.119) are positive. The final value of $\mathrm{H}(M_{t+1}|E_t)$ depends on the actual probabilities but the maximum value for n positive terms is $\log n$. So the more different co-action entities there are for a time t the higher the conditional entropy $\mathrm{H}(M_{t+1}|E_t)$ can get. Also note that if there are no actions at t i.e. no co-action entity in no co-action trajectory at t then $\mathrm{H}(M_{t+1}|E_t) = 0$. Entity actions of entities in $\mathfrak{E}^{PA}$ are therefore necessary and sufficient for $\mathrm{H}(M_{t+1}|E_t) > 0$.

The conditional entropy $\mathrm{H}(M_{t+1}|E_t)$ measures the uncertainty about the next agent state when the current environment state is known. It has been proposed as a measure of non-heteronomy in Bertschinger et al. (2008). Non-heteronomy means that the agent is not determined by the history of the environment. We only treat here the case where the history length is just one time-step E_t but generalisations to multiple time-steps $\mathrm{H}(M_{t+1}|E_{t-l:t})$ are straightforward (see remark to Definition 76). We have argued above that non-heteronomy at time t depends on the existence of entity actions performed at t and is limited by the number of different such actions at t. The entity actions as we have defined them are therefore like building blocks that make up the non-heteronomy of a stochastic process. In other words they are a local version of non-heteronomy in the case of value actions. Furthermore our definition applies to extent actions as well.

The measure of autonomy proposed by Bertschinger et al. (2008) contains another component measuring self-determination i.e. the degree to which the current agent state determines its next state. This is not ensured by our definition of actions. If M_{t+1} is an independently and uniformly distributed random variable then there are $|\mathcal{M}_{t+1}|$ co-actions at t and $\mathrm{H}(M_{t+1}|E_t) = \mathrm{H}(M_{t+1}) = \log|\mathcal{M}_{t+1}|$. The role of ensuring self-determination in our case is delegated to the entity set. The perception-action loop entity set $\mathfrak{E}^{PA}$ is too nonrestrictive for this purpose. Our notion of ι-entities on the other hand would not count a value m_{t+1} of an independently distributed random variable as part of a larger entity .

We expect that there are multiple action definitions that coincide with conditions similar to autonomy/non-heteronomy in the case of perception-action loop entities. An example would be requiring that the environments at $t+1$ differ in some way. We have only presented one definition here. Investigating and comparing further possible notions is future work.

We note that Ikegami and Taiji (1998) propose to use possible/compatible counterfactual trajectories of game players as signs of autonomy. This idea is similar to ours. We construct the capability to act from the counterfactual trajectories. Actions are arguably the basic units that realise autonomy over a longer period of time. In this sense we have in fact used the counterfactual trajectories to allow for a kind of autonomy.

5.6.2 Entity perception in the perception-action loop

We now look at how entity perception as defined in Section 5.5 specialises to the case of the perception-action loop. This argument in effect consti-

tutes a proof that our Definition 87 of the branch-morph is a generalisation of the conditional probability distributions $p_{M_{t+1}}(.|m_t, e_t) : \mathcal{M}_{t+1} \to [0,1]$ to non-interpenetrating, co-perception entities that (in contrast to the case of the perception-action loop) do not exhaust a set of future random variables and may exhibit counterfactual variation in extent. This result is not surprising since we set out to do just this but it is also instructive to work through the recovery of the original expression of the conditional probability distribution starting from the general branch-morph.

We pick an entity m_T from the entity set $\mathfrak{E}^{PA}$ and consider its perceptions at an arbitrary time-step $t \in T$. In order to get the perceptions at t we need

1. the co-perception entities $\mathfrak{S}(m_T, t)$ of m_T at t,
2. the branching partition $\eta(m_T, t)$ with its branches,
3. the co-perception environments,
4. the branch-morphs for each environment,
5. and the co-perception environment partition $\pi^{\mathfrak{S}}(x_A, t)$ with its blocks, the perceptions.

These can be identified in the following way.

1. The co-perception entities $\mathfrak{S}(m_T, t)$ are the entities in $\mathfrak{E}^{PA}$ that have non-empty time-slices at $t, t+1$, and that are identical to m_T up to t. All entities in $\mathfrak{E}^{PA}$ have non-empty time slices at all times. So we have:

$$\mathfrak{S}(m_T, t) = \{\bar{m}_T \in \mathfrak{E}^{PA} : \bar{m}_{\leq t} = m_{\leq t}\} \tag{5.122}$$

   Note that as t increases there are less and less co-perception entities. At $t = n-1$ (recall that $T = 0 : n-1$) we eventually have $\mathfrak{S}(m_T, t) = \{m_T\}$. Also note that the futures of the co-perception entities exhaust the future random variables $M_{t<}$ i.e.

$$\mathfrak{S}(m_T, t)_{t<} = \{\bar{m}_{t<} : \bar{m}_T \in \mathfrak{S}(m_T, t)\} = \mathcal{M}_{t<}. \tag{5.123}$$

2. First recall that the entity set $\mathfrak{E}^{PA}$ satisfies non-interpenetration since they all occupy the same set $\{M_t\}_{t\in T}$ of random variables. Therefore $\mathfrak{S}(m_t, t)$ is mutually exclusive and we get unique perception via the branching partition $\eta(m_T, t)$ of the entire set $\mathfrak{S}(m_t, t)$. The branching partition $\eta(m_T, t)$ is composed out of blocks (the branches) of co-perception entities that are identical up to $t+1$ i.e.

$$\begin{aligned} \hat{m}_T \sim \bar{m}_T \\ \Leftrightarrow \hat{m}_{t+1} = \bar{m}_{t+1}. \end{aligned} \tag{5.124}$$

   We can therefore identify the blocks of $\eta(m_T, t)$ i.e. the future branches by the values that the entities take at $t+1$. Define the branch $b(\bar{m}_{t+1})$ associated to $\bar{m}_{t+1} \in \mathcal{M}_{t+1}$ via

$$b(\bar{m}_{t+1}) := \{\hat{m}_T \in \mathfrak{S}(m_T, t) : \hat{m}_{t+1} = \bar{m}_{t+1}\}. \tag{5.125}$$

The branching partition is then:

$$\eta(m_T, t) = \{b(\bar{m}_{t+1}) \subseteq \mathfrak{S}(m_T, t) : \bar{m}_{t+1} \in \mathcal{M}_{t+1}\}. \tag{5.126}$$

3. The co-perception environments are the STPs $x_{V_t \setminus A_t}$ compatible with at least one co-perception entity. For the perception-action loop and entity m_T at t we have $\mathcal{X}_{V_t \setminus A_t} = \mathcal{E}_t$ and therefore $\mathcal{X}^{\mathfrak{S}}_{V_t \setminus A_t} = \mathcal{E}^{\mathfrak{S}}_t$. Where $\mathcal{E}^{\mathfrak{S}}_t$ is

$$\mathcal{E}^{\mathfrak{S}}_t = \{e_t \in \mathcal{E}_t : \exists \bar{m}_T \in \mathfrak{S}(m_T, t), p_{M_T, E_t}(\bar{m}_T, e_t) > 0\}. \tag{5.127}$$

As we have noted in Section 5.5.5 since the co-perception entities exhaust $M_{t<}$ this requirement is equivalent to

$$\mathcal{E}^{\mathfrak{S}}_t = \{e_t \in \mathcal{E}_t : p_{M_t, E_t}(\bar{m}_t, e_t) > 0\}. \tag{5.128}$$

4. The branch-morphs are the probability distributions $p_{\eta(m_T,t)}(.|e_t, m_{\leq t}) : \eta(m_T, t) \to [0, 1]$ over the branches for each co-perception environment $e_t \in \mathcal{E}^{\mathfrak{S}}_t$. These are defined using Eq. (5.114) which for the perception-loop becomes

$$p(b(\bar{m}_{t+1}), e_t | m_{\leq t}) := \sum_{\hat{m}_T \in b(\bar{m}_{t+1})} p_{M_{t<}, E_t}(\hat{m}_{t<}, e_t | m_{\leq t}). \tag{5.129}$$

We can rewrite the sum on the right hand side using Eq. (5.125) for $b(\bar{m}_{t+1})$ and then $\mathfrak{S}(m_T, t)_{t<} = \mathcal{M}_{t<}$:

$$\begin{aligned}
p(b(\bar{m}_{t+1}), e_t | m_{\leq t}) &= \sum_{\{\hat{m}_T \in \mathfrak{S}(m_T,t) : \hat{m}_{t+1} = \bar{m}_{t+1}\}} p_{M_{t<}, E_t}(\hat{m}_{t<}, e_t | m_{\leq t}) && (5.130)\\
&= \sum_{\{\hat{m}_{t<} \in \mathcal{M}_{t<} : \hat{m}_{t+1} = \bar{m}_{t+1}\}} p_{M_{t<}, E_t}(\hat{m}_{t<}, e_t | m_{\leq t}) && (5.131)\\
&= \sum_{\hat{m}_{t+1<} \in \mathcal{M}_{t+1<}} p_{M_{t<}, E_t}(\bar{m}_{t+1}, \hat{m}_{t+1<}, e_t | m_{\leq t}) && (5.132)\\
&= p_{M_{t+1}, E_t}(\bar{m}_{t+1}, e_t | m_{\leq t}). && (5.133)
\end{aligned}$$

The definition of the branch-morph for the perception-action loop is

$$p_{\eta(m_T,t)}(b(\bar{m}_{t+1}) | e_t, m_{\leq t}) := \frac{p(b(\bar{m}_{t+1}), e_t | m_{\leq t})}{\sum_{b \in \eta(m_T,t)} p(b, e_t | m_{\leq t})} \tag{5.134}$$

$$\tag{5.135}$$

which we can rewrite now

$$\begin{aligned}
p_{\eta(m_T,t)}(b(\bar{m}_{t+1}) | e_t, m_{\leq t}) &= \frac{p_{M_{t+1}, E_t}(\bar{m}_{t+1}, e_t | m_{\leq t})}{\sum_{\hat{m}_{t+1} \in \mathcal{M}_{t+1}} p_{M_{t+1}, E_t}(\hat{m}_{t+1}, e_t | m_{\leq t})} && (5.136)\\
&= \frac{p_{M_{t+1}, E_t}(\bar{m}_{t+1}, e_t | m_{\leq t})}{p_{E_t}(e_t | m_{\leq t})} && (5.137)\\
&= p_{M_{t+1}}(\bar{m}_{t+1} | e_t, m_{\leq t}) && (5.138)\\
&= p_{M_{t+1}}(\bar{m}_{t+1} | e_t, m_t). && (5.139)
\end{aligned}$$

In the last line we used the Bayesian network of the perception-action loop.

5. The co-perception environment partition $\pi^{\mathfrak{S}}(m_T, t)$ of $\mathcal{E}_t^{\mathfrak{S}} =$ is the partition induced by the equivalence classes of the equivalence relation

$$\begin{aligned} &\hat{e}_t \sim \bar{e}_t \\ &\quad \Leftrightarrow \forall b \in \eta(m_T, t) : p_{\eta(m_T,t)}(b|\hat{e}_t, m_{\leq t}) = p_{\eta(m_T,t)}(b|\bar{e}_t, m_{\leq t}). \end{aligned} \tag{5.140}$$

Using the branch-morph above this is equivalent to

$$\begin{aligned} &\hat{e}_t \sim \bar{e}_t \\ &\quad \Leftrightarrow \forall m_{t+1} \in \mathcal{M}_{t+1} : p_{M_{t+1}}(m_{t+1}|\hat{e}_t, m_t) = p_{M_{t+1}}(m_{t+1}|\bar{e}_t, m_t) \end{aligned} \tag{5.141}$$

which is just the equivalence relation of Eq. (3.71) used to extract the sensor-values in Section 3.3.6.

So we have seen that our definitions of Section 5.5 specialise in the case of the perception-action loop to the same concept of perception as in Section 3.3.6.

More interesting for future research is that the branch-morphs are generalisations of the conditional probability distribution $p_{M_{t+1}}(.|e_t, m_t)$. These conditional probability distributions play a role in various information theoretic concepts formulated for the perception-action loop. Examples include informational closure (Bertschinger et al., 2006), autonomy (Bertschinger et al., 2008), and morphological computation (Zahedi and Ay, 2013). We have also related entity actions to information theoretic measures in the perception action loop. The branch-morph (and also the entity actions) therefore suggest that it is possible to generalise these measures to entity sets that are non-exhaustive and vary counterfactually in extent. As we have argued in Section 5.3 such entities should be considered in a general definition of agents in multivariate Markov chains. The branch-morphs (and possibly other similar constructions) therefore provide a new tool to formulate such a general definition of agents.

Chapter 6

The structure of spatiotemporal patterns in small Markov chains

In this chapter we investigate the structure of integrated and completely locally integrated spatiotemporal patterns as it is revealed by the disintegration hierarchy. This will expose many of the formal notions introduced in Chapter 3 in practice. For this we will use two very simple multivariate Markov chains. We will use the disintegration theorem (Theorem 22) to extract the completely locally integrated spatiotemporal patterns. We will also use the SLI symmetry theorem and its corollary (Theorems 30 and 31) to explain the structure of the disintegration hierarchies.

In Section 5.3.5.2 we proposed to employ the completely locally integrated spatiotemporal patterns as a formal definition for entities. The entity set obtained in this way are called the ι-entities. We will calculate the entity sets for

We will see in Section 6.5 that the three phenomena that entity definitions should not preclude (compositionality, degree of freedom traversal, counterfactual variation) are exhibited by ι-entities in the example systems.

In Sections 5.4 and 5.5 we defined entity actions and entity perceptions for arbitrary entity-sets. In Section 6.6 we present examples of entity actions and entity perceptions of ι-entities in the example systems.

In Chapter 7 we discuss the results of this chapter, point out weaknesses, and propose further research directions. While there are some promising signs with respect to using ι-entities as entity sets and our notions of entity action and entity perception there are also some problems that need to be addressed before these proposals can claim to capture agents that only lack goal-directedness.

This chapter exclusively investigates examples of original notions that we have presented in Chapters 3 and 5. Therefore, almost its entire content is original. For reference the contributions of this chapter are:

- Computation and presentation of disintegration and refinement-free dis-

integration hierarchies for two simple systems.

- Explanation of the occurrence of multiple disconnected components in the partially ordered disintegration levels via the SLI symmetry theorems.
- Computation and presentation of the completely locally integrated spatiotemporal patterns of two simple systems.
- Examples of ι-entities that exhibit the three phenomena compositionality, degree of freedom traversal, and counterfactual variation that we argued for in Section 5.3.
- Examples of entity actions of ι-entities.
- Example of interpenetrating ι-entities showing that they do not necessarily obey non-interpenetration.
- Example of an entity perception and a branch-morph using a proxy for a co-perception partition.
- Example of an entity action and entity perception of the same ι-entity at the same time-step.
- Discussion of the results on ι-entities as entity sets in the example systems.

As we will see in Section 6.1 the computational complexity of computing the completely locally integrated spatiotemporal patterns increases rapidly. Since an important aspect of our proposal to use these patterns as entities is that they do not rely on further intuitions or prior knowledge about what entities are we must consider all possibilities exhaustively. The multivariate Markov chains we choose in this section are therefore extremely small. While this limits the relevance for the interpretation of the completely locally integrated spatiotemporal patterns as agents it allows us to discuss these patterns themselves more thoroughly.

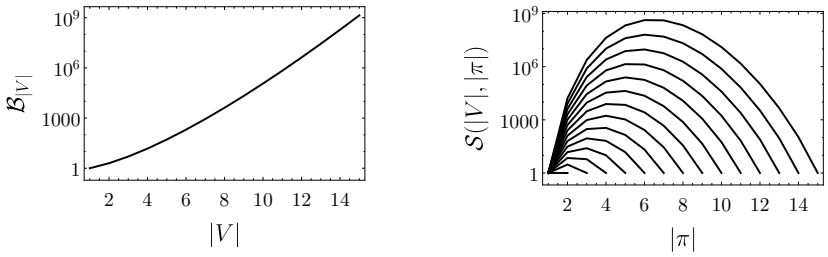
The systems we will look at are the following:

1. $MC^{=}$ which consists of two constant and independent binary random variables.
2. MC^{ϵ} which consists of two binary random variables that are mostly constant but where a noise term ϵ makes every other transition possible.

In all cases we choose a uniform initial distribution in order to exhaust the dynamics of all trajectories of the Markov chains.

6.1 Properties of partition lattices

Before we look at examples of partition lattices in the following chapters we quickly recall some properties that hold for all such lattices. As mentioned in



(a) Bell numbers

(b) Sterling numbers

Figure 6.1: Bell and Sterling numbers. (a) Logarithmic plot of the Bell numbers $\mathcal{B}_{|V|}$ for showing the number of partitions of a set V with $|V| = \{1, ..., 15\}$. (b) Logarithmic plot of the Sterling numbers showing the number of partitions π with $|\pi|$ blocks. The different lines correspond to different cardinalities $|V|$ of the set V that the π partition. The number at which a line ends indicates $|V|$.

Theorem 2 the number of elements $|\mathfrak{L}(V)|$ in a partition lattice $\mathfrak{L}(V)$ of a set V is given by the Bell number $\mathcal{B}_{|V|}$. Asymptotically, this number grows faster than exponentially with $|V|$ (De Bruijn, 1970) as is suggested by the more than linear growth in the logarithmic plot Fig. 6.1(a). We also mentioned that the number of partitions π with a fixed number $|\pi| = k$ of blocks is given by the Sterling number $\mathcal{S}(|V|, k)$. The most partitions then turn out to exist for the number of blocks k around $|V|/2$. This can be seen in Fig. 6.1(b).

6.2 Number of STP

The number $N_{pat}(\{X_i\}_{i\in V})$ of STPs in a Bayesian network $\{X_i\}_{i\in V}$ is equal to the the sum over all k of the subsets of size k times the number of different STP on this subset of size k. The number of different STP on a subset $A \subseteq V$ is $|\mathcal{X}_A|$ so we get:

$$N_{pat}(\{X_i\}_{i\in V}) = \sum_{A\subseteq V} |\mathcal{X}_A| = \sum_{k=1}^{|V|} \sum_{\{A\subseteq V:|A|=k\}} |\mathcal{X}_A|. \tag{6.1}$$

If we assume that the state spaces of all random variables in the network are equal i.e. $|\mathcal{X}_i| = n$ for all $i \in V$ then $|\mathcal{X}_A| = n^{|A|}$ and we get:

$$N_{pat}(\{X_i\}_{i\in V}) = \sum_{k=1}^{|V|} \binom{|V|}{k} n^k. \tag{6.2}$$

To get the number $N_{sli}(\{X_i\}_{i\in V})$ of SLI that have to be evaluated to check every partition of each STP x_A we have to further multiply the number of STPs $|\mathcal{X}_A|$ in Eq. (6.1) by the number of partitions $\mathcal{B}_{|A|}$ of these STPs. So the number of SLI values to evaluate is:

$$N_{sli}(\{X_i\}_{i\in V}) = \sum_{k=1}^{|V|} \sum_{\{A\subseteq V:|A|=k\}} |\mathcal{X}_A|\mathcal{B}_k. \tag{6.3}$$

For equal state spaces we get:

$$N_{sli}(\{X_i\}_{i\in V}) = \sum_{k=1}^{|V|} \binom{|V|}{k} n^k \mathcal{B}_k \tag{6.4}$$

SLI evaluations.

If we use the disintegration theorem we evaluate all partitions of the entire Bayesian network for each trajectory i.e.

$$N^{\mathfrak{D}}_{sli}(\{X_i\}_{i\in V}) = |\mathcal{X}_V|\mathcal{B}_{|V|} \tag{6.5}$$

partitions which is only the last term for $k = |V|$ of the sums over k in Eqs. (6.3) and (6.4). For equal state spaces this becomes

$$N^{\mathfrak{D}}_{sli}(\{X_i\}_{i\in V}) = n^{|V|}\mathcal{B}_{|V|}. \tag{6.6}$$

We still obtain what we are most interested in which are all the completely integrated patterns within the trajectories. However we also have to obtain the refinement free disintegration hierarchy which requires us to find the finest partitions at each disintegration level and check if they have refinements at preceding levels. If we ignore this for the moment, the disintegration theorem saves us

$$N_{sli}(\{X_i\}_{i\in V}) - N^{\mathfrak{D}}_{sli}(\{X_i\}_{i\in V}) = \sum_{k=1}^{|V|-1} \sum_{\{A\subseteq V:|A|=k\}} |\mathcal{X}_A|\mathcal{B}_k \tag{6.7}$$

evaluations. Which in the case of equal state spaces is:

$$N_{sli}(\{X_i\}_{i\in V}) - N^{\mathfrak{D}}_{sli}(\{X_i\}_{i\in V}) = \sum_{k=1}^{|V|-1} \binom{|V|}{k} n^k \mathcal{B}_k \tag{6.8}$$

evaluations. However as can be seen from Eqs. (6.5) and (6.6) the superexponential growth of the number of evaluations with the size of the index set V remains even if we use the disintegration theorem. The extra burden of finding the refinement free disintegration hierarchy has to be added to this as well.

We also note here that the above considerations do not include the computational resources needed to calculate the probabilities needed for the evaluation of specific local integrations. In order to calculate the disintegrations of a trajectory we need the global probability distribution over the entire Bayesian network. This means we need the probability (a real number between 0 and 1) of each trajectory. If we only have binary random variables, the number of trajectories is $2^{|V|}$ which make the straightforward computation of disintegration hierarchies unrealistic even for quite small systems. If we take a seven by seven grid of the game of life cellular automaton and want to look at three time-steps we have $|V| = 147$. If we use 32 bit floating numbers this give us 5×10^30 petabytes of storage needed for this probability distribution. This suggests that formal proofs are more useful for the investigation of specific local integration and disintegration hierarchies than simulations. Nonetheless we here show some very simple systems to get a better understanding of the formal notions.

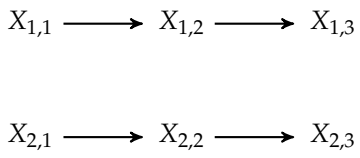


Figure 6.2: Bayesian network of $MC^{=}$. There is no interaction between the two processes.

6.3 Two constant and independent binary random variables: $MC^{=}$

6.3.1 Definition

Define the time- and space-homogeneous multivariate Markov chain $MC^{=}$ with Bayesian network $\{X_{j,t}\}_{j\in\{1,2\},t\in\{0,1,2\}}$ and

- $$\mathrm{pa}(j,t) = \begin{cases} \varnothing & \text{if } t = 0, \\ \{(j,t-1)\} & \text{else,} \end{cases} \tag{6.9}$$
- $$p_{j,t}(x_{j,t}|x_{j,t-1}) = \delta_{x_{j,t-1}}(x_{j,t}) = \begin{cases} 1 & \text{if } x_{j,t} = x_{j,t-1}, \\ 0 & \text{else,} \end{cases} \tag{6.10}$$
- $$p_{j,0}(x_{j,0}) = 1/4. \tag{6.11}$$

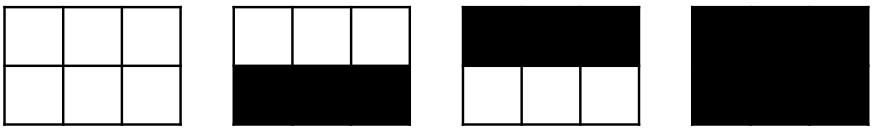
The Bayesian network can be seen in Fig. 6.2.

6.3.2 Trajectories

In order to get the disintegration hierarchy $\mathfrak{D}(x_V)$ we have to choose a trajectory x_V and calculate the SLI of each partition $\pi \in \mathfrak{L}(V)$. There are only four different trajectories possible in $MC^{=}$ and they are:

$$x_V = (x_{1,0}, x_{2,0}, x_{1,1}, x_{2,1}, x_{1,2}, x_{2,2}) = \begin{cases} (0,0,0,0,0,0) & \text{if } x_{1,0} = 0, x_{2,0} = 0; \\ (0,1,0,1,0,1) & \text{if } x_{1,0} = 0, x_{2,0} = 1; \\ (1,0,1,0,1,0) & \text{if } x_{1,0} = 1, x_{2,0} = 0; \\ (1,1,1,1,1,1) & \text{if } x_{1,0} = 1, x_{2,0} = 1. \end{cases} \tag{6.12}$$

Each of these trajectories has probability $p_V(x_V) = 1/4$ and all other trajectories have $p_V(x_V) = 0$. We call the four trajectories the *possible trajectories*. We visualise the possible trajectories as a grid with each cell corresponding to one variable. The spatial indices are constant across rows and time-slices V_t correspond to the columns. A white cell indicates a 0 and a black cell indicates a 1. This results in the grids of Fig. 6.3.



(a) $(0,0,0,0,0,0)$ *(b)* $(0,1,0,1,0,1)$ *(c)* $(1,0,1,0,1,0)$ *(d)* $(1,1,1,1,1,1)$

Figure 6.3: Visualisation of the four possible trajectories of $MC^{=}$. In each trajectory the time index increases goes from left to right. There are two rows corresponding to the two random variables at each time step and three columns corresponding to the three time-steps we are considering here.

6.3.3 Partitions of trajectories

The disintegration hierarchy is composed out of all partitions in the lattice of partitions $\mathfrak{L}(V)$. Note that we are partitioning the entire spatially and temporally extended index set V of the Bayesian network and not only the time-slices. Blocks in the partitions of $\mathfrak{L}(V)$ are then, in general, spatiotemporal patterns and not only spatial patterns.

The number of partitions $|\mathfrak{L}(V)|$ of a set of $|V| = 6$ elements is $\mathcal{B}_6 = 203$ (see Theorem 2). These partitions π can be classified according to their cardinality $|\pi|$ (number of blocks in the partition). The number of partitions of a set of cardinality $|V|$ into $|\pi|$ blocks is the Sterling number $\mathcal{S}(|V|, |\pi|)$. For $|V| = 6$ we find the Sterling numbers:

$\lvert\pi\rvert$	1	2	3	4	5	6
$\mathcal{S}(\lvert V\rvert, \lvert\pi\rvert)$	1	31	90	65	15	1

(6.13)

It is important to note that the partition lattice $\mathfrak{L}(V)$ is the same for all trajectories as it is composed out of partitions of V. On the other hand the values of SLI $\mathrm{mi}_\pi(x_V)$ with respect to the partitions in $\mathfrak{L}(V)$ generally depend on the trajectory x_V.

6.3.4 SLI values of the partitions

We can calculate the SLI $\mathrm{mi}_\pi(x_V)$ of every trajectory x_V with respect to each partition $\pi \in \mathfrak{L}(V)$ according to Definition 49:

$$\mathrm{mi}_\pi(x_V) := \log \frac{p_V(x_V)}{\prod_{b\in\pi} p_b(x_b)}. \tag{6.14}$$

In the case of $MC^{=}$ the SLI values with respect to each partition do not depend on the trajectories. For an overview we plotted the values of SLI with respect to each partition $\pi \in \mathfrak{L}(V)$ for any trajectory of $MC^{=}$ in Fig. 6.4. We can see in Fig. 6.4 that the cardinality does not determine the value of SLI. At the same time there seems to be a trend to higher values of SLI with increasing cardinality of the partition. We can also observe that only five different values of SLI are attained by partitions on this trajectory. We will collect these classes of partitions with equal SLI values in the disintegration hierarchy next.

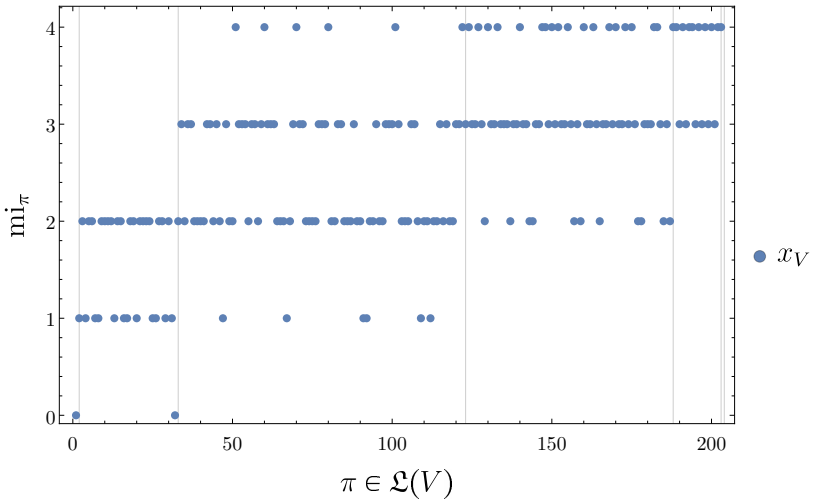


Figure 6.4: Specific local integrations $\mathrm{mi}_\pi(x_V)$ *of any of the four trajectories* x_V *seen in Fig. 6.3 with respect to all* $\pi \in \mathfrak{L}(V)$*. The partitions are ordered according to an enumeration with increasing cardinality* $|\pi|$ *(see Pemmaraju and Skiena, 2009, chap. 4.3.3 for the method). We indicate with vertical lines at what partitions the cardinality* $|\pi|$ *increases by one.*

6.3.5 Disintegration hierarchy

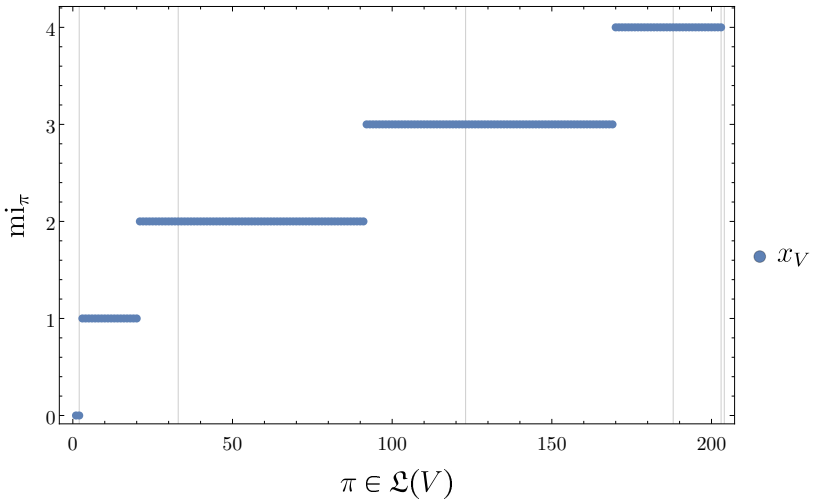


Figure 6.5: Same as Fig. 6.4 but with the partitions sorted according to increasing SLI.

In order to get insight into the internal structure of the partitions of a trajectory x_V we obtain the disintegration hierarchy $\mathfrak{D}(x_V)$ (see Definition 54) look at the Hasse diagrams of each of the disintegration levels $\mathfrak{D}_i(x_V)$. If we sort the partitions of any trajectory of $MC^=$ according to increasing SLI value we obtain Fig. 6.5. There we see groups of partitions attaining the SLI values

$\{0, 1, 2, 3, 4\}$ these groups are the disintegration levels $\{\mathfrak{D}_1(x_V), \mathfrak{D}_2(x_V), \mathfrak{D}_3(x_V), \mathfrak{D}_4(x_V), \mathfrak{D}_5(x_V)\}$. The exact numbers of partitions in each of the levels are:

i	1	2	3	4	5
mi_π	0	1	2	3	4
$\lvert\mathfrak{D}_i\rvert$	2	18	71	78	34

(6.15)

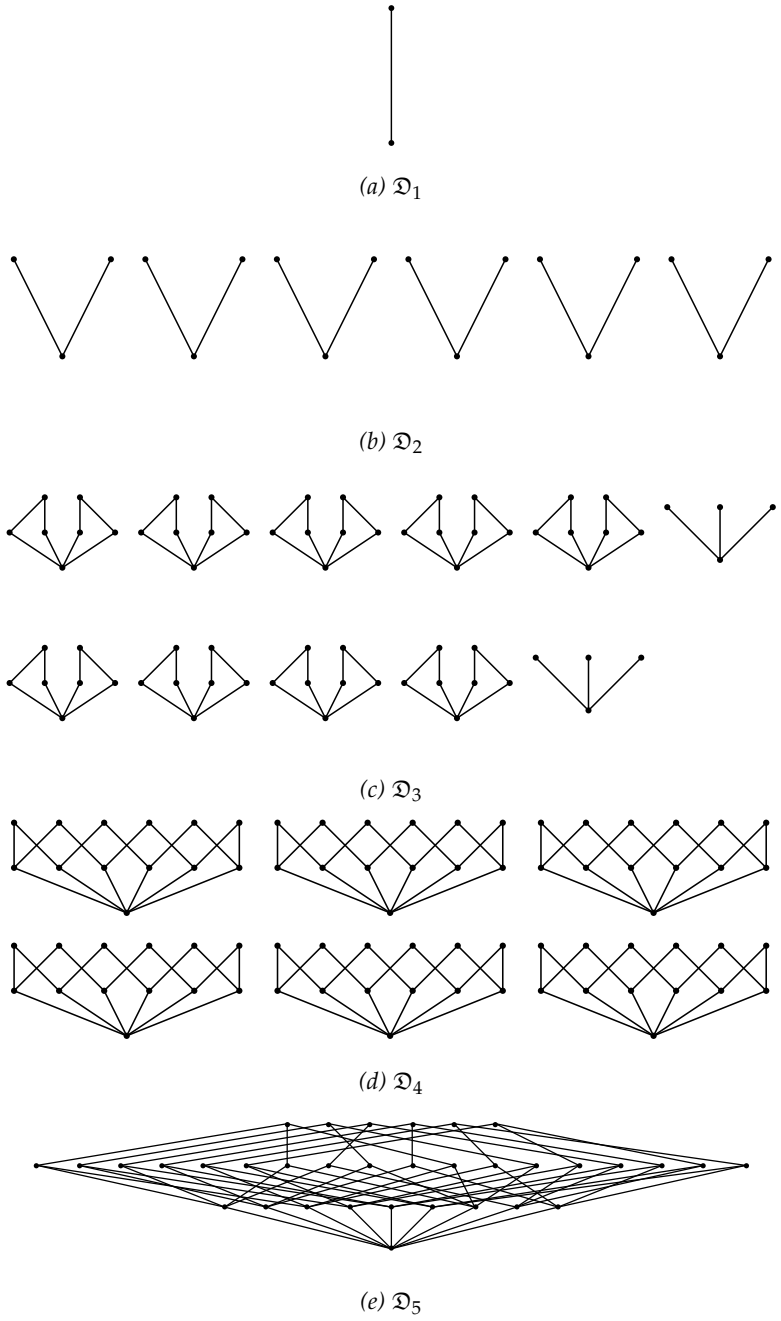
Next we look at the Hasse diagram of each of those disintegration levels. Since the disintegration levels are subsets of the partition lattice $\mathfrak{L}(V)$ they are in general not lattices by themselves. The Hasse diagrams visualise the set of partitions in each disintegration level partially ordered by refinement $\vartriangleleft$ (see Definition 13). Recall that in Hasse diagrams of such posets the partitions are arranged such that if $\pi \neq \xi$ and $\pi \vartriangleleft \xi$ then π is drawn below ξ. Also, an edge is drawn from partition π to ξ if one covers the other e.g. if $\pi \vartriangleleft: \xi$. The Hasse diagrams are shown in Fig. 6.6. We see immediately that within each disintegration level apart from the first and the last the Hasse diagrams contain multiple connected components.

Furthermore, within a disintegration level the connected components often have the same Hasse diagrams. For example in $\mathfrak{D}_2$ (Fig. 6.6(b)) we find six connected components with three partitions each. The identical refinement structure of the connected components is related to the symmetries of the probability distribution over the trajectories. This will be discussed in Section 6.3.6. We can visualise the partitions themselves in the Hasse diagrams as in Fig. 6.7.

Recall that due to the disintegration theorem (Theorem 22) we are interested especially in partitions that do not have refinements at their own or any preceding (i.e. lower indexed) disintegration level. These partitions consist of blocks that are completely integrated i.e. all possible partitions of each of the blocks results in a positive SLI value or is a single node of the Bayesian network. The refinement-free disintegration hierarchy $\mathfrak{D}^{\blacktriangleleft}(x_V)$ contains only these partitions and is shown in a Hasse diagram in Fig. 6.8.

6.3.6 Symmetries

As shown in Theorem 36.(i) the symmetries of the trajectory x_V that are also symmetries of p_V generate partitions with respect to which the SLI has the same value. More formally, if for all elements h of a group of permutations $\mathfrak{H}$ we have both $hp_V = p_V$ and $hx_V = x_V$ (for the particular trajectory x_V, not necessarily for all trajectories) then $\mathrm{mi}_{h\pi}(x_V) = \mathrm{mi}_\pi(x_V)$. This means if we start with one partition π with a particular SLI value (i.e. on a particular disintegration level) then we can generate the orbit of partitions $\mathfrak{H}\pi := \{h\pi : h \in \mathfrak{H}\}$ under $\mathfrak{H}$ which contains only partitions with equal SLI value. Since any permutation preserves the refinement relation between and cardinality of partitions (see Theorems 25 and 26 respectively) we can also take a set of partitions partially ordered by refinement and generate identical posets (with identical Hasse diagrams) of partitions of respectively equal cardinality. In



(a) $\mathfrak{D}_1$

(b) $\mathfrak{D}_2$

(c) $\mathfrak{D}_3$

(d) $\mathfrak{D}_4$

(e) $\mathfrak{D}_5$

Figure 6.6: Hasse diagrams of the five disintegration levels of the trajectories of $MC^{=}$*. Every vertex corresponds to a partition and edges indicate that the lower partition refines the higher one.*

the following we will find the symmetries of both p_V and of the four possible trajectories.

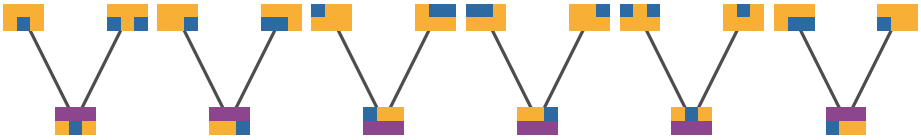


Figure 6.7: Hasse diagram of $\mathfrak{D}_2$ of $MC^{=}$ trajectories. Here we visualise the partitions at each vertex. The blocks of a partition are the cells of equal colour. Note that we can obtain all six disconnected components by permuting the indices via spatial inversion $h_{\updownarrow}(j,t) = (|J| + 1 - j, t)$ and "global" time shifts $h_{\rightarrow}(j,t) = (j, (t-1) \mod 3)$. For example acting on the partitions in the first component from the left we obtain: the second component via $h_{\rightarrow}$, the third component via $h_{\updownarrow} \circ h_{\rightarrow}^{-1}$, the fourth via $h_{\updownarrow} \circ h_{\rightarrow}$, the fifth via $h_{\updownarrow}$, and the sixth via $h_{\rightarrow}^{-1}$.

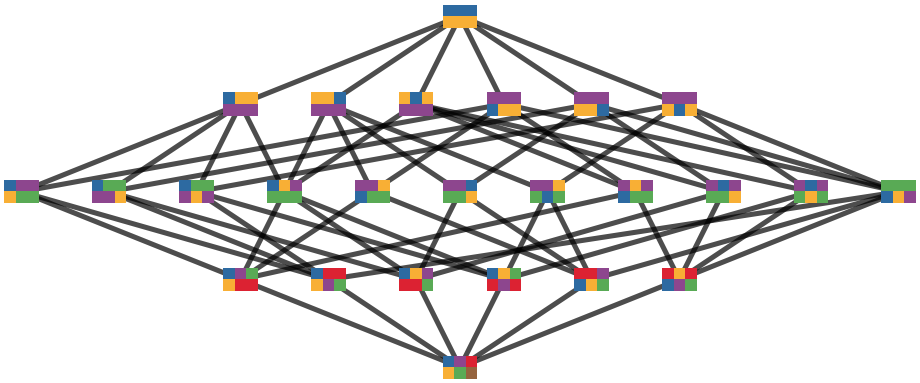


Figure 6.8: Hasse diagrams of the refinement-free disintegration hierarchy $\mathfrak{D}^{\blacktriangleleft}$ of $MC^{=}$ trajectories. Here we visualise the partitions at each vertex. The blocks of a partition are the cells of equal colour. It turns out that partitions that are on the same horizontal level in this diagram correspond exactly to a level in the refinement-free disintegration hierarchy $\mathfrak{D}^{\blacktriangleleft}$. The i-th horizontal level starting from the top corresponds to $\mathfrak{D}_i^{\blacktriangleleft}$. Take for example the second horizontal level from the top. The partitions on this level are just the minimal elements of the poset $\mathfrak{D}_2$ which was visualised in Fig. 6.7. We have shown the posets of the other disintegration levels only without their partitions in Fig. 6.6 but their minimal elements are all present at the according horizontal level in this diagram.

We will now establish symmetries of p_V. It is not difficult to just look at the probability distribution p_V which attributes the probability $1/4$ for each of the four possible trajectories and infer symmetries by visual inspection. We only need to find permutations of the index set V that transform all trajectories into trajectories with the same probability. Since all four possible trajectories in Fig. 6.3 have the same probability every permutation that maps these trajectories onto each other is a symmetry of p_V. Note that permutations which map one of those trajectories into a trajectory with probability zero cannot be symmetries of p_V. To get an intuition for this take one of the less symmetric possible trajectories of $MC^{=}$ from Fig. 6.3 e.g. the second one

Any permutation of the cells in this grid is a permutation of V. Since the other possible trajectories are either more symmetric (like the first and the fourth) or share the same symmetry (like the third) any permutation which maps this trajectory to one of the four is a symmetry of p_V. Such symmetries are "spatial inversion" (flipping the top and bottom row in Fig. 6.3), "global time-slice permutations" (reordering the columns in Fig. 6.3 in any way) , and "individual time-slice permutation" (reordering the cells within a single row in Fig. 6.3 in any way). Spatial inversion maps the second trajectory above to the third and vice versa, the first and second trajectory are left invariant. The time-slice permutations both leave all possible trajectories invariant. Note that the global time-slice permutations can actually be constructed from the individual time-slice permutations. An example of a permutation that is not a symmetry is to invert only a single time-slice and not change the others (flipping the top and bottom cell in the second column). Next, we will formally derive theses intuitions from the definition of $MC^{=}$.

First let us look at "individual time-slice permutations" more precisely this means permutations of the form $h(j,t_1) = (j,t_2)$ where for different spatial indices the time-shift may be different i.e. for $j_1 \neq j_2$ we may have $h(j_1,t_1) = (j_1,t_2)$, $h(j_2,t_1) = (j_2,t_3)$, and $t_2 \neq t_3$. We can separate such permutations into two sets, those that permute the time indices only for the first spatial index and those that only permute them for the second. Let $\mathfrak{T}_j = \Sigma_{(j,T)} \subset \Sigma_V$ be the subgroup of permutations only affecting the indices in $(j,T) = \{(j,t) : t \in T\}$ (we will in the following loosely refer to $(1,T),(2,T)$ as the first and second, or top and bottom *row* since this is visually intuitive given our representation of trajectories and patterns as grids). Note that by combining elements $h_1 \in \mathfrak{T}_1$ and $h_2 \in \mathfrak{T}_2$ to get $h_1 \circ h_2$ we can affect both $(1,T)$ and $(2,T)$ so that we can also permute whole time-slices at once if both h_1 and h_2 permute the respective sets in the same way. This gives the "global time-slice permutations".

Now note that for the present case both $\mathfrak{T}_j$, $j \in \{1,2\}$ are actually symmetry groups of all possible trajectories x_V. This means they are necessarily symmetry groups of p_V since for any arbitrary p_V if $x_V^g = x_V$ then of course $gp_V(x_V) = p_V(x_V^g) = p_V(x_V)$. Formally, for any $j \in \{1,2\}$ and $h_j \in \mathfrak{T}_j$ we have

$$
\begin{aligned}
h_j x_V = h_j x_{J,T} &= (x_{i,t})_{i\in J,t\in T}^{h_j^{-1}} && (6.16)\\
&= \{X_{i,t} = x_{i,t}\}_{i\in J,t\in T}^{h_j^{-1}} && (6.17)\\
&= \{X_{i,t} = x_{h_j^{-1}(i,t)}\}_{i\in J,t\in T} && (6.18)\\
&= \{X_{i,t} = x_{i,t}\}_{i\neq j,t\in T} \cup \{X_{i,t} = x_{h_j^{-1}(j,t)}\}_{j,t\in T} && (6.19)
\end{aligned}
$$

If we now recall that for all possible trajectories x_V and $t,s \in T$ we have

$x_{j,t} = x_{k,s}$ so that $x_{h_j^{-1}(j,t)} = x_{j,t}$ we see that

$$h_j x_V = x_V \tag{6.20}$$

for all possible trajectories. Since all permutations are bijective, the impossible trajectories must also be mapped to impossible trajectories such that all $h_j \in \mathfrak{T}_j$ with $j \in \{1,2\}$ are symmetries of all trajectories and therefore also symmetries of p_V.

Next we will look at spatial inversion. Note first that spatial inversion does not leave all the possible trajectories invariant since it transforms the second into the third and vice versa. So spatial inversion is only a symmetry of p_V.

It would not be difficult to derive that spatial inversion is a symmetry of p_V directly from looking at its effect on p_V. However, we here want to exhibit how Theorem 33 can be used in establishing symmetries.

Theorem 33 tells us that a group of spatial symmetries of both the Markov matrix and the initial distribution is also a symmetry group of p_V. Define the spatial inversion via $h_\updownarrow(j,t) = (|J| + 1 - j, t)$. Then $h_\updownarrow \circ h_\updownarrow = \mathrm{i}$, $h_\updownarrow^{-1} = h_\updownarrow$ which means $\{h_\updownarrow, \mathrm{i}\}$ form a subgroup of the spatial permutations $\Sigma_J \times \{\mathrm{i}\}$. We now show that this is a symmetry group of p_V. Note that $V_0 = (J,0) = \{(1,0),(2,0)\}$, and recall that we chose the uniform distribution as initial distribution such that for any $x_{V_0}, \bar{x}_{V_0} \in \mathcal{X}_{V_0}$ we have $p_{V_0}(x_{V_0}) = p_{V_0}(\bar{x}_{V_0}) = 1/4$. Since by construction $x_{V_0}^{h_\updownarrow} \in \mathcal{X}_{V_0}$ we have for any $x_{V_0} \in \mathcal{X}_{V_0} = \{(0,0),(0,1),(1,0),(1,1)\}$:

$$\begin{aligned}
h_\updownarrow p_{V_0}(x_{V_0}) &= p_{V_0}(x_{V_0}^{h_\updownarrow}) && (6.21)\\
&= \Pr(\{X_{1,0} = x_{1,0}, X_{2,0} = x_{2,0}\}^{h_\updownarrow}) && (6.22)\\
&= \Pr(X_{1,0} = x_{h_\updownarrow(1,0)}, X_{2,0} = x_{h_\updownarrow(2,0)}) && (6.23)\\
&= \Pr(X_{1,0} = x_{2,0}, X_{2,0} = x_{1,0}) && (6.24)\\
&= p_{V_0}(x_{V_0}). && (6.25)
\end{aligned}$$

Or in short $h_\updownarrow p_{V_0} = p_{V_0}$. We also have $h_\updownarrow(P p_{V_t}) = P(h_\updownarrow p_{V_t})$. To see this recall the definition of the dynamics of each of the two random variables of $MC^=$ in Eq. (6.10) which was:

$$p_{j,t}(x_{j,t}|x_{j,t-1}) = \delta_{x_{j,t-1}}(x_{j,t}) = \begin{cases} 1 & \text{if } x_{j,t} = x_{j,t-1}, \\ 0 & \text{else.} \end{cases} \qquad \text{(6.10 revisited)}$$

From this we get the Markov matrix via

$$\begin{aligned}
p_{V_{t+1}}(x_{V_{t+1}}|x_{V_t}) &= \prod_{i \in V_{t+1}} p_i(x_i|x_{\mathrm{pa}(i)}) && (6.26)\\
&= p_{1,t}(x_{1,t+1}|x_{1,t}) p_{2,t}(x_{2,t+1}|x_{2,t}) && (6.27)\\
&= \delta_{x_{1,t}}(x_{1,t+1}) \delta_{x_{2,t}}(x_{2,t+1}). && (6.28)
\end{aligned}$$

So that (being extra verbose as it is the only such calculation in this thesis)

$$p_{V_{t+1}}(x_{V_{t+1}}^{h_\updownarrow}|x_{V_t}^{h_\updownarrow}) = \Pr(\{X_{j,t+1} = x_{j,t+1}\}_{j\in J}^{h_\updownarrow}|\{X_{j,t} = x_{j,t}\}_{j\in J}^{h_\updownarrow}) \tag{6.29}$$

$$\begin{aligned} &= \Pr((X_{1,t+1} = x_{1,t+1})^{h_\updownarrow}|(X_{1,t} = x_{1,t})^{h_\updownarrow}) \\ &\qquad \times \Pr((X_{2,t+1} = x_{2,t+1})^{h_\updownarrow}|(X_{2,t} = x_{2,t})^{h_\updownarrow}) \end{aligned} \tag{6.30}$$

$$\begin{aligned} &= \Pr(X_{1,t+1} = x_{h_\updownarrow(1,t+1)}|X_{1,t} = x_{h_\updownarrow(1,t)}) \\ &\qquad \times \Pr(X_{2,t+1} = x_{h_\updownarrow(2,t+1)}|X_{2,t} = x_{h_\updownarrow(2,t)}) \end{aligned} \tag{6.31}$$

$$\begin{aligned} &= \Pr(X_{1,t+1} = x_{2,t+1}|X_{1,t} = x_{2,t}) \\ &\qquad \times \Pr(X_{2,t+1} = x_{1,t+1}|X_{2,t} = x_{1,t}) \end{aligned} \tag{6.32}$$

$$= p_{1,t+1}(x_{2,t+1}|x_{2,t})p_{2,t+1}(x_{1,t+1}|x_{1,t}) \tag{6.33}$$

$$= \delta_{x_{2,t}}(x_{2,t+1})\delta_{x_{1,t}}(x_{1,t+1}) \tag{6.34}$$

$$= \delta_{x_{1,t}}(x_{1,t+1})\delta_{x_{2,t}}(x_{2,t+1}) \tag{6.35}$$

$$= p_{V_{t+1}}(x_{V_{t+1}}|x_{V_t}). \tag{6.36}$$

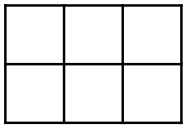
which implies

$$p_V(x_{J,t+1}|x_{J,t}) = p_V(x_{J,t+1}^{h_\updownarrow}|x_{J,t}^{h_\updownarrow}). \tag{6.37}$$

By Theorem 36.(ii) this implies $h_\updownarrow(Pp_{V_t}) = P(h_\updownarrow p_{V_t})$ and together with $h_\updownarrow p_{V_0} = p_{V_0}$ Theorem 33 then implies $h_\updownarrow p_V = p_V$ which shows that $h_\updownarrow$ is a symmetry of p_V. Clearly, $\{\mathrm{i}, h_\updownarrow\}$ is then a symmetry group of p_V.

We have now established that $\mathfrak{T}_1, \mathfrak{T}_2$, and $\{\mathrm{i}, h_\updownarrow\}$ are symmetry groups of p_V. Since they are all subgroups of Σ_V we can combine their elements via function composition (e.g. $h_j \circ h_\updownarrow$) to get (possibly further) elements of Σ_V. The set of elements that can be formed in this way is a subgroup of Σ_V called the subgroup generated by $\mathfrak{T}_1, \mathfrak{T}_2$, and $\{\mathrm{i}, h_\updownarrow\}$. Since all elements of this subgroup are also symmetries of p_V we call the subgroup generated by $\mathfrak{T}_1, \mathfrak{T}_2$, and $\{\mathrm{i}, h_\updownarrow\}$ the symmetry group of $MC^=$ and denote it by $\mathfrak{G}_{MC^=}$.

Let us now come back to the symmetries of SLI. If we look at the first trajectory



we see that it is symmetric with respect to any permutation of V. So every symmetry of p_V is also a symmetry of x_V and we can use the entire group $\mathfrak{G}_{MC^=}$ of symmetries of p_V to generate equally disintegrating partitions. Formally, for each $g \in \mathfrak{G}_{MC^=}$ we have $gx_V = x_V$ so that $\mathrm{mi}_{g\pi}(x_V) = \mathrm{mi}_\pi(x_V)$. As we have seen in Fig. 6.7 combining $h_\updownarrow \in \mathfrak{G}_{MC^=}$ and $h_\rightarrow \in \mathfrak{G}_{MC^=}$ (a time shift to the right of all indices) generates all the disconnected components of the poset of $\mathfrak{D}_2$.

If we look at the other disintegration levels we can see that $\mathfrak{G}_{MC^=}$ explains the occurrence of multiple disconnected components also in $\mathfrak{D}_4$. It is compatible also with $\mathfrak{D}_1$ in the sense that for the two partitions in $\mathfrak{D}_1$ (see Fig. 6.10(a))

we have $g\pi = \pi$ for all $g \in \mathfrak{G}_{MC^=}$. Similarly we only expect a single component in $\mathfrak{D}_5$ since the minimal partition there, the zero of $\mathfrak{L}(V)$, also has $g\mathbf{0} = \mathbf{0}$ for all $g \in \mathfrak{G}_{MC^=}$.

The only outlier in this hierarchy is $\mathfrak{D}_3$. We show the poset of partitions of $\mathfrak{D}_3$ in Fig. 6.9. It shows two different kinds of disconnected component. One consisting of seven partitions which occurs nine times and one consisting of four partitions occurring two times. According to Theorems 25 and 26 we should not be able to use symmetries to map partitions in the first kind of component to partitions in the second kind.

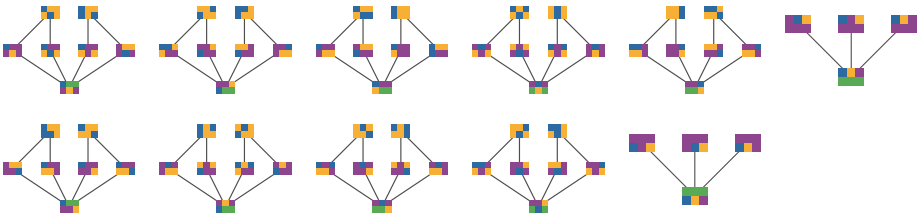


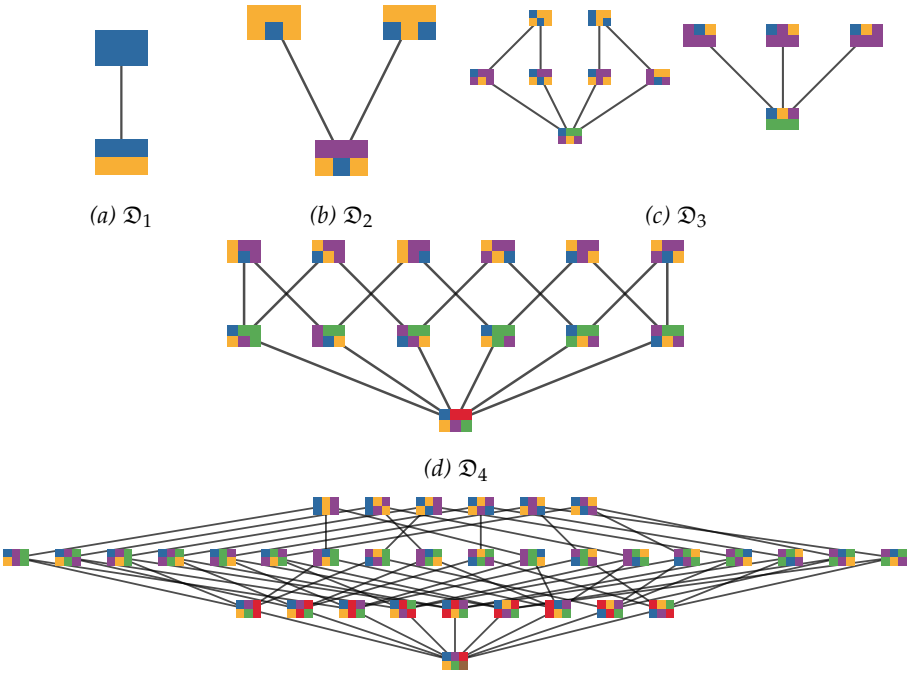
Figure 6.9: Hasse diagram of $\mathfrak{D}_3$ of $MC^=$ trajectories with visualised partitions. The blocks of a partition are the cells of equal colour. Note that we can obtain all nine disconnected components with three horizontal levels by permuting the indices via spatial inversion $h_\updownarrow(j,t) = (|J|+1-j,t)$ and "individual time slice permutations" permuting indices in one row independent of the other. However, we cannot obtain the partitions of the two components (at the right edge of both rows) with two horizontal levels from the partitions of the components with three horizontal level components. The equality of SLI

This can be seen by noting for example that in most cases the cardinalities of the blocks of the partitions in the components of the first kind are different from the cardinalities of the blocks in the components of the second kind. In case of the first kind the cardinalities are: top layer: $\{4,2\}$ and $\{3,3\}$, second layer: $\{4,1,1\},\{3,2,1\},\{2,2,2\}$, and bottom layer $\{2,2,1,1\}$. In the the case of the second kind: top layer: $\{4,1,1\}$, bottom layer: $\{3,1,1,1\}$. Since all permutations in Σ_V maintain the cardinalities of all blocks only the partitions with block cardinalities $\{4,1,1\}$ occur in either kind and could be transformed into each other by an elements of $\mathfrak{G}_{MC^=}$. However, we can observe that this is impossible as we only have $h_\updownarrow$ which exchanges the entire rows and cannot exchange individual cells in the top row with those in the bottom row.

So the symmetries of p_V due to permutations of V do not explain why the components of the first and second kind occur on the same disintegration level i.e. why the partitions they contain have the same SLI values. This is then a kind of degeneracy of the disintegration level $\mathfrak{D}_3$.

In order to get an overview of only the different components of the posets in each disintegration level we have drawn them including visualised partitions in Fig. 6.10.

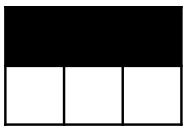
Finally let us look at the second trajectory again



(a) $\mathfrak{D}_1$ *(b)* $\mathfrak{D}_2$ *(c)* $\mathfrak{D}_3$ *(d)* $\mathfrak{D}_4$ *(e)* $\mathfrak{D}_5$

Figure 6.10: Representatives of the distinct Hasse diagrams within each disintegration level of $MC^{=}$*. Again we visualise the partitions at each vertex with the blocks of a partition of equal colour. Note that at level* $\mathfrak{D}_3$ *(in (c)) there are two distinct Hasse diagrams whereas on the other levels there is only one per level.*

and apply $h_{\updownarrow}$ we get



which is the third trajectory. So $h_{\updownarrow}$ is not a symmetry of the second trajectory and we don't have $h_{\updownarrow}x_V = x_V$. This means that the conditions for Theorem 36.(i) are not satisfied. Nonetheless, as we mentioned before, the disintegration hierarchies of all four possible trajectories are identical. The reason for this is Theorem 36.(ii) which asserts that even if $gx_V \neq x_V$ we still have $\mathrm{mi}_{g\pi}(x_V) = \mathrm{mi}_{\pi}(x_V)$ if for all blocks $b \in \pi$ we have $p_b(x_b^g) = p_b(x_b)$. This is in fact the case here for $g \in \mathfrak{G}_{MC^{=}}$ and all partitions on the second and third trajectories.

To see this note first that for x_V either the second or third possible trajectory any $b \subseteq V$, $x_b^g = (x_{g(i)})_{i \in b}$ is still a pattern that can occur on at least one of the four possible trajectories. In other words, it is impossible that x_b^g is a pattern that cannot occur. Also keep in mind that x_b^g concerns exactly the same random variables in the Bayesian network as x_b even if the values it fixes may differ. Then distinguish two situations. First, let $b \subset (j, T)$ i.e. b is part of a single row indicated by j. Then x_b^g fixes some of the values in the one of the rows and x_b also fixes some of the values (possibly different) in one of the rows (possibly the other one). Recall that only four trajectories are possible and fixing the value (independent of what value) of any random variable in one of the rows selects two possible trajectories from the four. Then $p_b(x_b^g)$ is a sum of the probabilities of these two trajectories and since their probabilities are all $1/4$ this sum is $1/2$ for both $p_b(x_b^g)$ and $p_b(x_b)$. Second, let $b \cap (1, T) \neq \varnothing$ and $b \cap (2, T) \neq \varnothing$ then x_b^g and also x_b fixes the values in both rows so there is only a single possible trajectory selected whose probability is $1/4$ in every case. So again $p_b(x_b^g) = p_b(x_b)$.

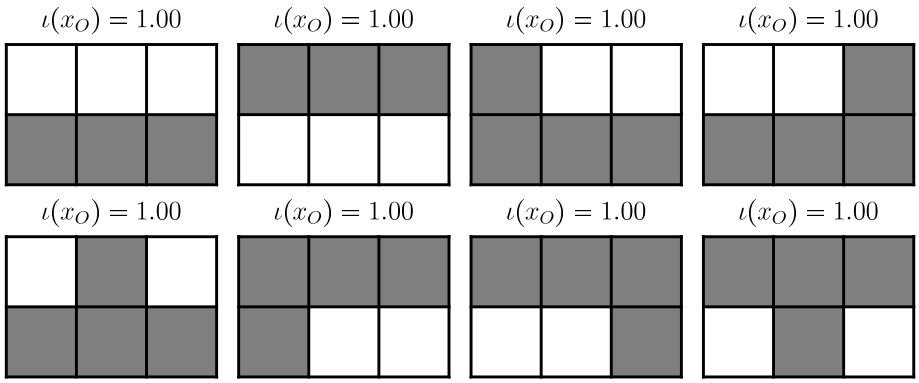


Figure 6.11: All distinct completely integrated STPs on the first possible trajectory of $MC^=$. The value of complete local integration is indicated above each STP. We display STPs by colouring the cells corresponding to random variables that are not fixed to any value by the STP in grey. Cells corresponding to random variables that are fixed by the STP are coloured according to the value i.e. white for 0 and black for 1.

6.3.7 Completely integrated STPs

After having looked at the disintegration hierarchy extensively we now make use of it by extracting the completely (locally[1]) integrated STPs of the four trajectories of $MC^=$. Recall that due to the disintegration theorem (Theorem 22) we know that all blocks in partitions that occur in the refinement-free disintegration hierarchy are either singletons or correspond to completely integrated STPs. If we look at the refinement-free disintegration hierarchy in Fig. 6.8 we

[1]When it is clear from context that we are talking about complete local integration we drop "local" for the sake of readability.

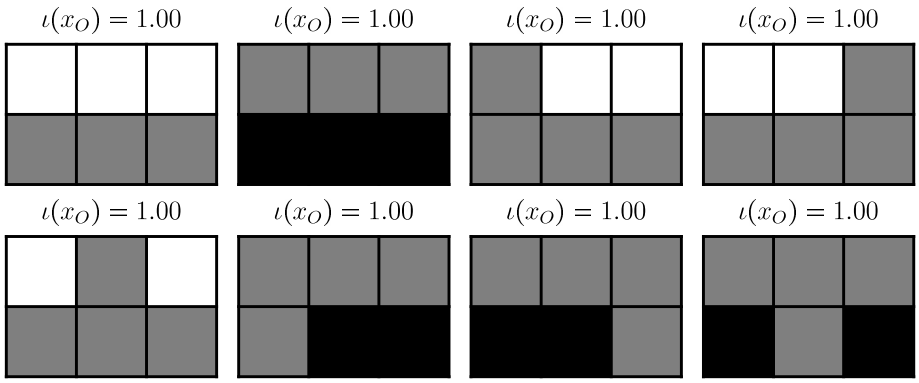


Figure 6.12: All distinct completely integrated STPs on the second possible trajectory of $MC^{=}$. The value of complete local integration is indicated above each STP.

see that many blocks occur in multiple partitions and across disintegration levels. We also see that there are multiple blocks that are singletons. If we ignore singletons since they are trivially integrated as they cannot be partitioned we end up with eight different blocks. Since the disintegration hierarchy is the same for all possible trajectories these blocks are also the same for each of them. However, the STPs that result are different due to the different values within the blocks. We show the eight completely integrated STPs and their complete local integration (Definition 53) on the first trajectory in Fig. 6.11 and on the second trajectory in Fig. 6.12.

Since the disintegration hierarchies are the same for the four possible trajectories of $MC^{=}$ we get the same refinement-free partitions and therefore the same blocks containing the completely integrated STPs. This is apparent when comparing Figs. 6.11 and 6.12 and noting that each STP occurring on the first trajectory has a corresponding STP on the second trajectory that differs (if at all) only in the values of the cells it fixes and not in what values it fixes. More visually speaking, for each STP in Fig. 6.11 there is a corresponding STP in Fig. 6.11 leaving the same cells grey.

If we are not interested in a particular trajectory we can also look at all different completely integrated STP on any trajectory. For $MC^{=}$ these are shown in Fig. 6.13 We see that all completely integrated STPs x_O have the same value of complete local integration $\iota(x_O) = 1$. This can be explained using the deterministic expression for the SLI of Eq. (4.10) and noting that for $MC^{=}$ if any of the values $x_{j,t}$ is fixed by a STP then $(x_{j,s})_{s\in T} = x_{j,T}$ are determined since they must be the same value. This means that the number of trajectories $N(x_{j,S})$ in which any pattern $x_{j,S}$ with $S \subseteq T$ occurs is either $N(x_{j,S}) = 0$, if the pattern is impossible, or $N(x_{j,S}) = 2$ since there are two trajectories compatible with it. Note that all blocks x_b in any of the completely integrated STP and all STP x_O themselves are of the form $x_{j,S}$ with $S \subseteq T$. Let

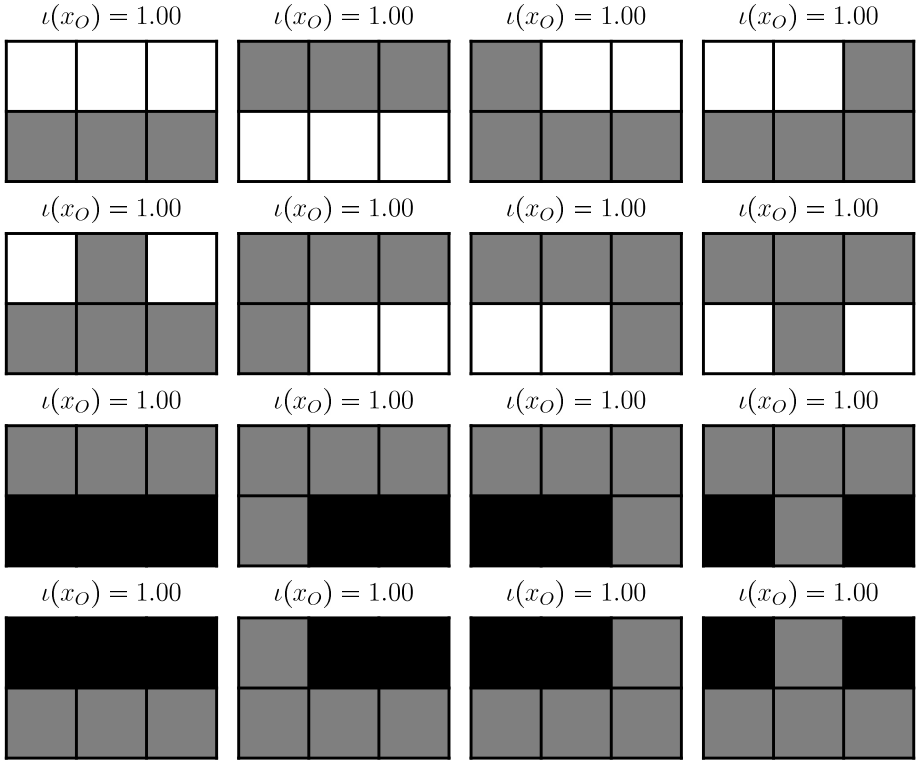


Figure 6.13: All distinct completely integrated STPs on all four possible trajectories of $MC^{=}$. The value of complete local integration is indicated above each STP.

$N(x_{j,S}) =: N$ and plug this into Eq. (4.10) for an arbitrary partition π:

$$\mathrm{mi}_\pi(x_O) = (|\pi| - 1)\log|\mathcal{X}_{V_0}| - \log\frac{\prod_{b\in\pi} N(x_b)}{N(x_O)} \tag{6.38}$$

$$= (|\pi| - 1)\log|\mathcal{X}_{V_0}| - \log\frac{N^{|\pi|}}{N} \tag{6.39}$$

$$= (|\pi| - 1)\log\frac{|\mathcal{X}_{V_0}|}{N}. \tag{6.40}$$

To get the complete local integration value we have to minimise this with respect to π where $|\pi| \geqslant 2$. So for $|\mathcal{X}_{V_0}| = 4$ and $N = 2$ we get $\iota(x_O) = 1$.

Another observation is that the completely integrated STPs are all limited to one of the two rows. This shows on a simple example that, as we would expect, completely integrated patterns cannot extend from one independent process to another.

6.4 Two random variables with small interactions

In this section we look at a system almost identical to that of Section 6.3 but with a kind of noise introduced. This allows all trajectories to occur and is

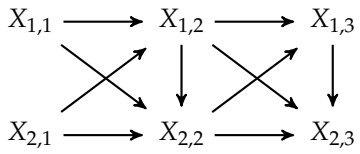


Figure 6.14: Bayesian network of MC^{ϵ}.

designed to test whether the spatiotemporal patterns maintain integration in the face of noise.

6.4.1 Definition

We define the time- and space-homogeneous multivariate Markov chain MC^{ϵ} via the Markov matrix P with entries

$$P_{f(x_{1,t+1},x_{2,t+1}),f(x_{1,t},x_{2,t})} = p_{J,t+1}(x_{1,t+1}, x_{2,t+1}|x_{1,t}, x_{2,t}) \tag{6.41}$$

where we define the function $f : \{0,1\}^2 \to [1:4]$ via

$$f(0,0) = 1, f(0,1) = 2, f(1,0) = 3, f(1,1) = 4. \tag{6.42}$$

With this convention P is

$$P = \begin{pmatrix} 1-3\epsilon & \epsilon & \epsilon & \epsilon \\ \epsilon & 1-3\epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & 1-3\epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & 1-3\epsilon \end{pmatrix} \tag{6.43}$$

The initial distribution is again the uniform distribution

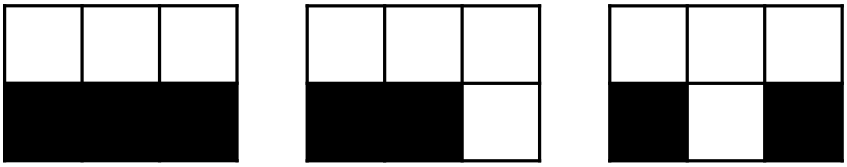
$$p_{j,0}(x_{j,0}) = 1/4. \tag{6.44}$$

Writing this multivariate Markov chain as a Bayesian network is possible but the conversion is tedious. The Bayesian network one obtains can be seen in Fig. 6.14. The state of *both* random variables remains the same with probability $1-3\epsilon$ and transitions into each other possible combination with probability ϵ. In the following we set $\epsilon = 1/100$.

6.4.2 Trajectories

In this system all trajectories are possible trajectories. This means there are $2^6 = 64$ possible trajectories, since every one of the six random variables can be in any of its two states. There are three classes of trajectories with equal probability of occurring. The first class with the highest probability of occurring are the four possible trajectories of $MC^{=}$. Then there are 24 trajectories that make a single ϵ-transition (i.e. a transition where the next pair is not the

same as the current one $(x_{1,t+1}, x_{2,t+1}) \neq (x_{1,t}, x_{2,t})$, these transitions occur with probability ϵ), and 36 trajectories with two ϵ-transitions. We pick only one trajectory from each class. The representative trajectories are shown in Fig. 6.15 and will be denoted x_V^1, x_V^2, and x_V^3 respectively. The probabilities are $p_V(x_V^1) = 0.235225, p_V(x_V^2) = 0.0024250, p_V(x_V^3) = 0.000025$.



(a) $x_V^1 = (0,1,0,1,0,1)$ *(b)* $x_V^2 = (0,1,0,1,0,0)$ *(c)* $x_V^3 = (0,1,0,0,0,1)$

Figure 6.15: Visualisation of three trajectories of MC^ϵ. In each trajectory the time index increases from left to right. There are two rows corresponding to the two random variables at each time step and three columns corresponding to the three time-steps we are considering here. We can see that the first trajectory (in (a)) makes no ϵ-transitions, the second (in (b)) makes one from $t = 2$ to $t = 1$, and the third (in (c)) makes two.

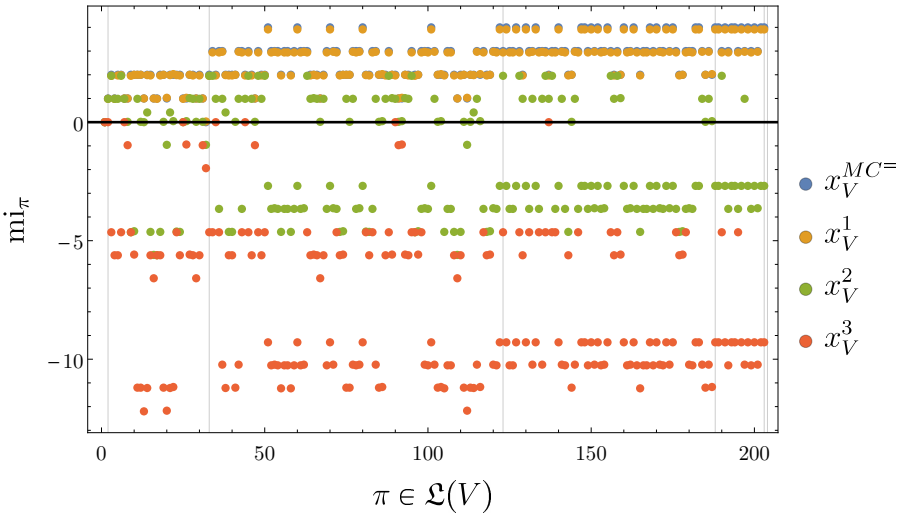


Figure 6.16: Specific local integrations $\mathrm{mi}_\pi(x_V)$ of one of the four trajectories of $MC^=$ (measured w.r.t. the probability distribution of $MC^=$), here denoted $x_V^{MC^=}$, and the three representative trajectories $x_V^k, x \in \{1,2,3\}$ of MC^ϵ (measured w.r.t. the probability distribution of MC^ϵ) seen in Fig. 6.15 with respect to all $\pi \in \mathfrak{L}(V)$. The partitions are ordered as in Fig. 6.4 with increasing cardinality $|\pi|$. Vertical lines indicate partitions where the cardinality $|\pi|$ increases by one. Note that the values of $x_V^{MC^=}$ are almost completely hidden from view by those of x_V^1.

6.4.3 SLI values of the partitions

Again we calculate the SLI $\mathrm{mi}_\pi(x_V)$ of every trajectory x_V with respect to each partition $\pi \in \mathfrak{L}(V)$. In contrast to $MC^=$ the SLI values with respect to each partition of MC^ϵ do depend on the trajectories. We plot the values of SLI with respect to each partition $\pi \in \mathfrak{L}(V)$ for the three representative trajectories in Fig. 6.16.

It turns out that the SLI values of x_V^1 are almost the same as those of $MC^=$ in Fig. 6.4 with small deviations due to the noise. This should be expected as x_V^1 is also a possible trajectory. Also note that trajectories x_V^2, x_V^3 exhibit negative SLI with respect to some partitions. In particular, x_V^3 has non-positive SLI values with respect to any partition. This is due to the low probability of this trajectory compared to its parts. The blocks of any partition have so much higher probability than the entire trajectory that the product of their probabilities is still greater or equal to the trajectory probability.

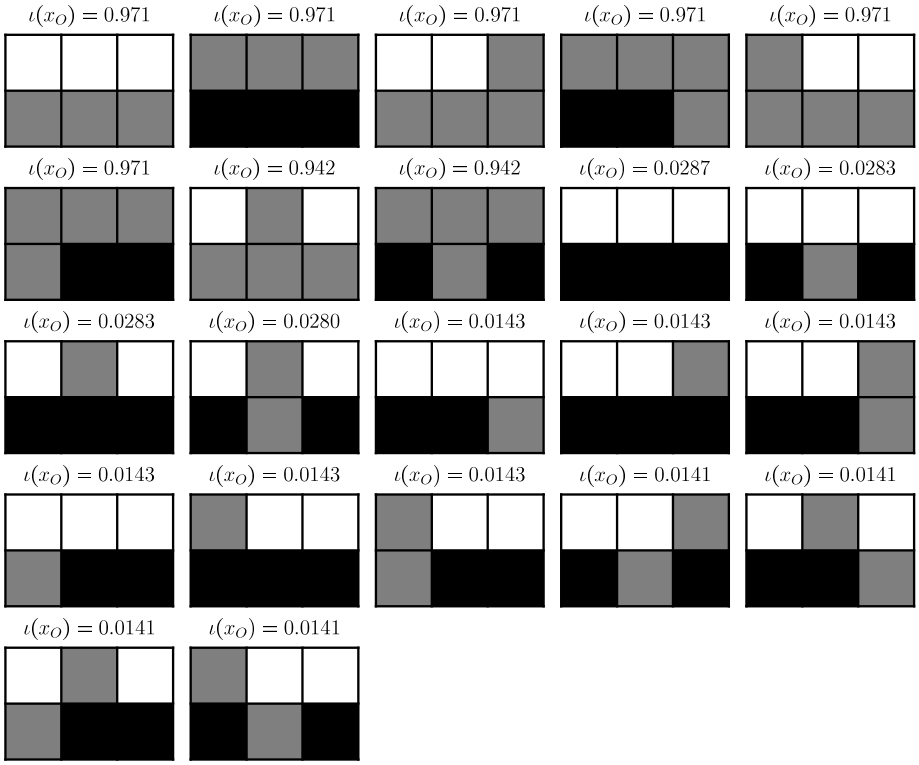


Figure 6.17: All distinct completely integrated STPs on the first trajectory x_V^1 of MC^ϵ. The value of complete local integration is indicated above each STP. See Fig. 6.11 for colouring conventions.

6.4.4 Completely integrated STPs

In this section we look at the completely integrated STPs for each of the three representative trajectories $x_V^k, k \in \{1,2,3\}$. They are visualised together with their complete local integration values in Figs. 6.17 to 6.19. In contrast to the situation of $MC^=$ we now have completely integrated STPs with varying values of complete local integration.

On the first trajectory x_V^1 we find all the eight STPs that are completely locally integrated in $MC^=$ (see Fig. 6.12). These are also more than an order of magnitude more integrated than the rest of the completely integrated STPs.

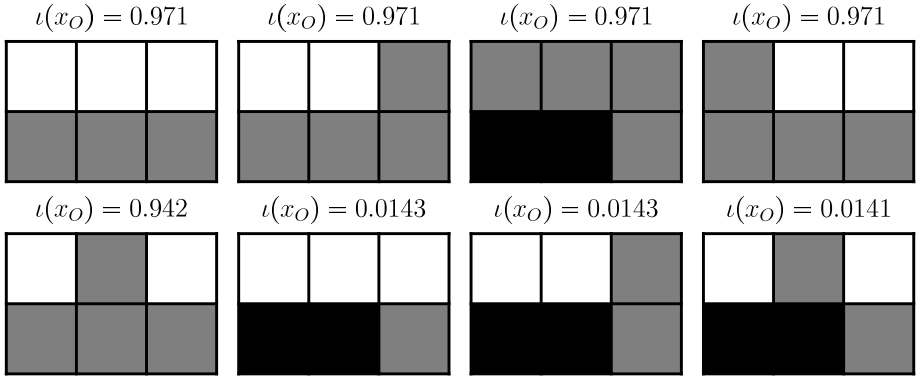


Figure 6.18: All distinct completely integrated STPs on the second trajectory x_V^2 of MC^ϵ. The value of complete local integration is indicated above each STP.

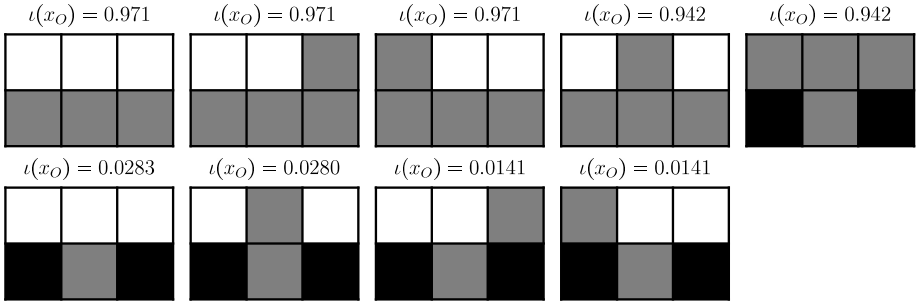


Figure 6.19: All distinct completely integrated STPs on the third trajectory x_V^3 of MC^ϵ. The value of complete local integration is indicated above each STP.

6.5 Completely locally integrated spatiotemporal patterns as entities

In Section 5.3.5 we have proposed to use the set of completely integrated spatiotemporal patterns to solve the identity problem. This means using the com-

pletely integrated spatiotemporal patterns as an entity set. There we called this entity set the ι-entities. In this section we look at the completely integrated spatiotemporal patterns found in Sections 6.3 and 6.4 with regard to the phenomena and properties of entities described in Section 5.3.

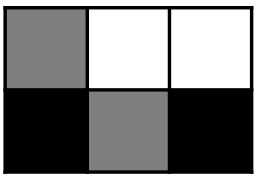
In Section 5.3 we have described three phenomena that should not be precluded by a formal notion of entities. These are

1. compositionality,
2. degree of freedom traversal, and
3. counterfactual variation.

We show by example that ι-entities can exhibit these three phenomena.

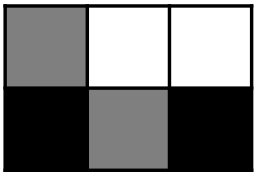
Regarding compositionality we can see in Fig. 6.13 that $MC^{=}$ contains ι-entities that are composed of multiple temporal parts. More precisely $MC^{=}$ contains ι-entities that occupy random variables in multiple time-slices. For example the ι-entity:

While $MC^{=}$ contains no ι-entities that are composed of multiple spatial parts such ι-entities exist in MC^{ϵ} on all three kinds of trajectories as we can see in Figs. 6.17 to 6.19. An example would be the ι-entity:



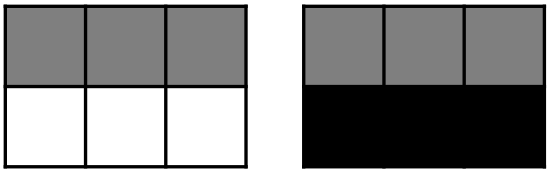
This entity is also temporally composite. These entities still have much lower ι values than for the entities that occupy only one of the random variables in multiple time-slices. However, this still shows that there are spatially, temporally, and spatiotemporally composite ι-entities.

Regarding degree of freedom traversal, we can see that, for example, the ι-entity



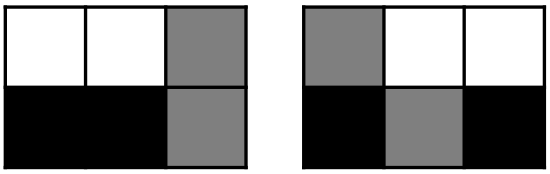
changes the degrees of freedom that it occupies from the bottom one to the top one and then to both in the two time-steps. So there are degree of freedom traversing ι-entities.[2]

Regarding counterfactual variation, the ι-entities are counterfactual *in value only* for $MC^{=}$. For example the following two ι-entities from the first and second possible trajectory differ in value on the same set of of occupied random variables:



But the sets of random variables occupied by the ι-entities of $MC^{=}$ are the same in all trajectories.

In MC^{ϵ} on the other hand we find that the sets of random variables occupied by the ι-entities differ from one trajectory to another. For example the ι-entities



which occur in x_V^2 and x_V^3 respectively occupy sets of random variables that no entity in the other trajectory occupies. So there the entity set of ι-entities for MC^{ϵ} is counterfactual in extent.

We only note briefly that there are some counter intuitive ι-entities that skip the second times-step for both $MC^{=}$ and MC^{ϵ} Figs. 6.17 to 6.19. Whether these are due to the small scale of the system or do occur more generally will be further investigated in the future.

6.6 Action and perception of ι-entities

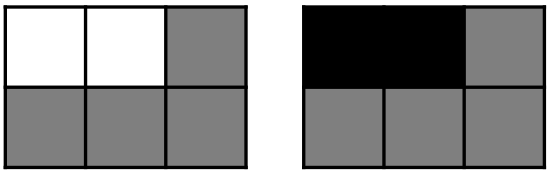
In this section we briefly present examples of ι-entities that exhibit actions and perceptions according to our definitions. We prove by example that that ι-entities can perform value and extent actions. We also show that non-interpenetration is not necessarily satisfied by ι-entities as it is not satisfied in our example systems $MC^{=}$ and MC^{ϵ}. This implies that the co-perception entities are not necessarily mutually exclusive and in fact we find example co-perception entities that are not mutually exclusive. We can still use a subset of mutually exclusive entities with common past to extract some perceptions. These are not uniquely defined however and we also choose a different subset

[2]Since there are also entities that do not traverse degrees of freedom in the first time-step we can also not just rename the indices at the second time-step to get rid of all degree of freedom traversing entities.

and show it has a slightly different branch-morph as well. We also show that the same ι-entity can perform an action and perceive something in the same transition from one step to the next.

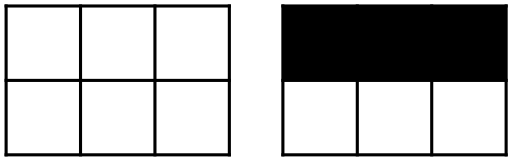
6.6.1 Actions of ι-entities in $MC^{=}$ and MC^{ϵ}

According to our definition of entity action (Definition 76) there are actions performed by ι-entities in $MC^{=}$ and MC^{ϵ}. For example the ι-entities



are co-action entities at the first time-step and their second time-step time-slices are co-actions.

- They occur in the different and possible co-action trajectories:

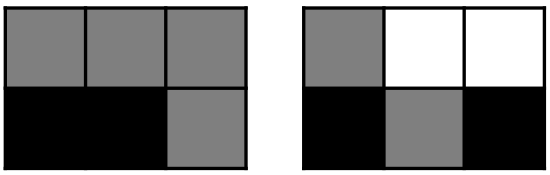


- They occupy the same random variable at the first time-step.
- The environments at the first time-step in the two trajectories are identical.
- And at the second time-step the two ι-entities differ.

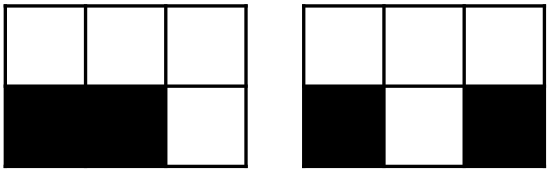
For the environment it is therefore impossible to be sure about the next configuration of the entity that co-occurs at the first time-step.

Since the above spatiotemporal patterns are ι-entities in the shown trajectories in both $MC^{=}$ and MC^{ϵ} they are co-action entities in both chains.

Note that the above is a value action. The time-slices at the second time-step differ only in the value they assign to the top random variable. In $MC^{=}$ there are no extent actions, but in MC^{ϵ} there are. The following two entities are co-action entities at the first time-step in co-action trajectories x^2_V and x^3_V of MC^{ϵ}:



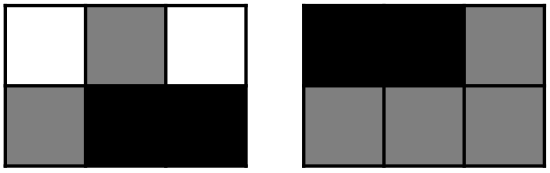
These entities occur in the different and possible co-action trajectories x^2_V and x^3_V:



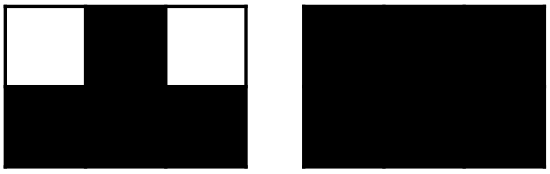
and also fulfil the other conditions for actions mentioned above. The action is an extent action since the occupied variable is the bottom variable for the left entity and the top variable for the right entity. This shows that there are ι-entities that perform value actions and those that perform extent actions.

The extent action above also reveals a challenge to ι-entities. We argued in Section 5.4 that the differences at $t+1$ of two entities with the same environments at t should be due to the entity or random. In the above case the entities (and their environments in the according trajectories) are equal at t so the differences between the two at $t+1$ are due to the noise. We suggested that it is the task of the entity-set to exclude such random parts from entities. The ι-entities might therefore need further adaptations or the notion of actions might need to be changed. A third possibility is that due to the "global" construction of entities there are some non-intuitive effects. Another such effect can be seen are the entities that skip an entire times-step that we mentioned in Section 6.5.

Finally, we note that there are also more intuitive co-action entities that perform extent actions. For example the co-action entities



which occur in the co-action trajectories

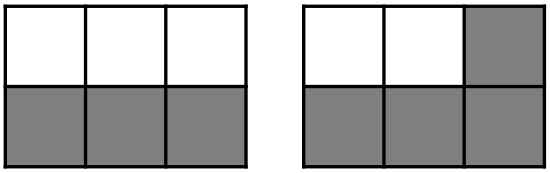


have different "internal" values at the first time-step which can be interpreted as the reason for the different extent at the second time-step. In the light of the previous extent action however this interpretation is questionable and needs further investigation.

6.6.2 Perceptions of ι-entities in $MC^=$ and MC^ϵ

Regarding perception we first note that the example chains $MC^=$ and MC^ϵ allow interpenetration of ι-entities. We can see that there are interpenetrating

ι-entities in all trajectories of the systems considered here. For example the ι-entities



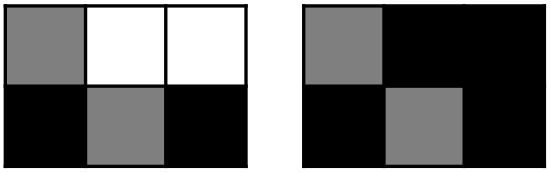
both occur in each of the three trajectories of MC^{ϵ}. They are not identical patterns and still both fix partly the same random variables. Yet, as they occur in the same trajectories the probability that they co-occur is non-zero. This violates non-interpenetration as defined in Definition 84. Therefore the dynamics of the systems considered here do not support the definition of perception proposed in Section 5.5. This implies that by itself the choice of the ι-entities as an entity set does not necessarily lead to non-interpenetration. As mentioned before it is not clear at what level non-interpenetration should be required.

Next, note that in $MC^{=}$ there are no co-perception entities for any of its ι-entities (see Fig. 6.13). For any given entity x_A with non-empty time-slices at t and $t+1$ there is no other entity y_B that has identical past $y_{B_{\leq t}} = x_{A_{\leq t}}$ and differs at $t+1$. This is due to the fact that the future of any entity is completely determined by its current state regardless of the rest of the system / environment.

In MC^{ϵ} we do find co-perception entities. Due to the interpenetration of entities however we cannot use the full set of co-perception entities. As mentioned in Section 5.5.8 we can use a mutually exclusive subset $\zeta(x_A, t)$ of the co-perception entities as a proxy for the co-perception partition. The simplest case is to use only two entities, the original entity x_A and a co-perception entity y_B so that $\zeta(x_A, t) = \{x_A, y_B\}$. The co-perception entity y_B must be chosen such that

- $\Pr(X_A = x_A, X_B = y_B) = 0$,
- $x_{A_{t+1}} \neq y_{B_{t+1}}$.

The first condition so that it is mutually exclusive the second so that it is part of a different branch in the branching partition. This gives us a branching partition $\eta(x_A, t) = \{\{x_a\}, \{y_B\}\}$. An example of such a co-perception pair are the following ι-entities of MC^{ϵ}:



So here $A = \{(2,0),(1,1),(1,2),(2,2)\}$ with $x_A = (1,0,0,1)$ and $B = A$ with $y_B = (1,1,1,1)$. Since $x_{A_{\leq t}} = y_{B_{\leq t}}$ and $A_0, A_1 \neq \varnothing$ we indeed have two

co-perception entities $y_B \in \mathfrak{S}(x_A, 0)$. The two co-perception entities perceive the difference between two co-perception environments at the first time-step $t = 0$. Since every trajectory is possible in MC^ϵ every environment at the first time-step is a co-perception environment so that $\mathcal{X}^{\mathfrak{S}}_{V_t \setminus A_t} = \mathcal{X}_{V_t \setminus A_t}$. The environments that will be classified are then the possible values $\{0, 1\}$ of the random variable $X_{V_t \setminus A_t} = X_{1,0}$ that is not fixed by the two co-perception entities.

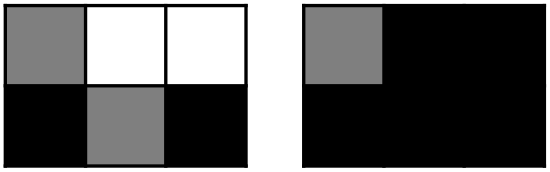
Each environment $x_{1,0} \in \mathcal{X}_{1,0}$ has an associated branch-morph $p(.|x_{1,0}, x_{A_{\preceq t}}) : \eta(x_A, t) \to [0, 1]$ over the two entities. The two morphs turn out to be:

	x_A	y_B
$p(.\|0, x_{A_{\preceq t}})$	4705/4754	49/4754
$p(.\|1, x_{A_{\preceq t}})$	49/4754	4705/4754

(6.45)

We see that they are indeed different such that the branch-morph partitions $\pi^{\mathfrak{S}}$ consists of two blocks containing one co-perception environment each $\pi^{\mathfrak{S}} = \{\{X_{1,0} = 0\}, \{X_{1,0} = 1\}\}$. The elements of this partition are then the perceptions.

Finally, we can see here that the choice of a co-perception entity is not unique. We can also choose an entity $z_C \neq y_B$ to get another co-perception pair. Another such possibility with the same x_A is the following pair:



Here $C = \{(2, 0), (1, 1), (2, 1), (1, 2), (2, 2)\}$ and $z_C = (1, 1, 1, 1, 1)$. Note that z_C and y_B differ at the second time-slice but are not mutually exclusive:

$$Pr(X_B = y_B, X_C = z_C) = 4753/20000 > 0. \tag{6.46}$$

Therefore, they cannot be put into the same branch (because they differ at $t + 1$) but they also cannot be put into different branches (because they can occur together). This is possible because they interpenetrate and makes it difficult to find a unique co-perception partition.

The perceptions of the co-perception pair x_A, z_C are the same as for x_A, y_B but this is due to the limits of our example. The branch morphs are different:

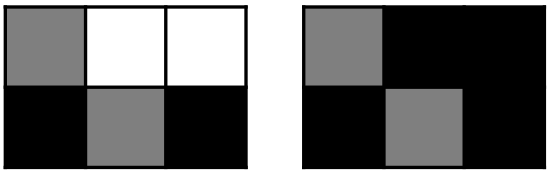
	x_A	z_C
$p(.\|0, x_{A_{\preceq t}})$	9410/9507	97/9507
$p(.\|1, x_{A_{\preceq t}})$	98/9507	9409/9507

(6.47)

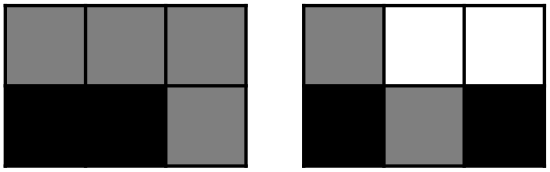
This indicates that in general the perceptions of different choices of $\zeta(x_A, t)$ co-perception entities are also different.

6.6.3 Action and perception of the same ι-entity at the same time

In the previous sections we have seen co-action entities and co-perception entities. We only want to emphasise here that an entity can have a co-perception pair and a (different) co-action pair at the same time t. This means that action and perception do not preclude each other. An example is the entity x_A from our perception example. It is the entity on the left in



shown together with its co-perception entity for the first time-step. However, x_A also occurs as a co-action at the first time-step of another entity. These co-actions (with x_A on the right) are



So the same ι-entity can perform actions and have perceptions the same time-step.

6.7 Discussion

In this chapter we presented two very simple multivariate Markov chains $MC^{=}$ and MC^{ϵ}. In Section 6.3 We calculated the disintegration hierarchy and the refinement free disintegration hierarchy of $MC^{=}$. We explained the occurrence of multiple identical disconnected components in the partially ordered disintegration levels. These are due to the invariance of the specific local integration as revealed by Theorems 30 and 31. We then presented the set of completely locally integrated spatiotemporal patterns of $MC^{=}$. This corresponds to the ι-entities of $MC^{=}$. These exhibit compositionality in time but not in space and counterfactual variation in value but not in extend. They do not exhibit degree of freedom traversal as should be expected due to the independent dynamics. All the entities of $MC^{=}$ have the same ι value of 1 bit. This is due to the four possible trajectories having identical probabilities and the deterministic dynamics.

In Section 6.4 we extracted the ι-entities of MC^{ϵ}. These exhibit compositionality in time and space, counterfactual variation in value and extend, and degree of freedom traversal. The ι-entities of MC^{ϵ} also attain various ι values ranging from 0.014 bit to 0.971 bit.

In Section 6.6 we turned our attention to entity action and entity perception of the ι-entities of $MC^{=}$ and MC^{ϵ}. There are entity actions in $MC^{=}$ but only in value. For our weak definition of actions this is not surprising since the entities in $MC^{=}$ are self-determining and independent of the rest of the system.

In MC^{ϵ} we also find entity actions in extend. Here we also find an example of an extent action that suggests that either the entity action requirements are too weak or the ι-entities do not provide the internal connection between their parts that we expected. We will discuss this further below.

Concerning perception, we find no perceptions in $MC^{=}$ which should be expected for parts of an independent process. In our formalisation of perception this leads to an empty set of co-perception entities. For MC^{ϵ} there are co-perception entities. However, we cannot use the unique construction of a branching partition due to interpenetration of ι-entities and non-exclusion of co-perception entities. We therefore used the approach that only relies on a set ζ of mutually exclusive co-perception entities to define the branch-morph. This resulted in perception being defined for entities in MC^{ϵ}. We also saw that another choice of a co-perception entity can lead to a (quantitatively) different perception. This further confirmed that without non-interpenetration perception is not necessarily uniquely defined.

Finally, we showed that the same ι-entity can perform entity actions and entity perception at the same time-step t. This is achieved simply by finding a co-action entity and a (usually different) co-perception entity for the same original entity at the same time-step. The story of perceptions and actions of an entity is then defined via the co-action and co-perception entities along its time-evolution.

The combination of ι-entities with entity action and entity perception then fulfils quite a few expectations and requirements that we have discussed before. However, there are also some things that are not easily interpreted. One thing briefly mentioned before are entities that skip a time-step. Such entities occur in both systems $MC^{=}$ and MC^{ϵ}. A second thing is the extent action that seemed to be random even though it was part of an entity. This could mean at least four different things:

1. The entity action definition is too weak and should require explicitly that the time-slice x_{A_t} of the entity at t determines to some degree the time-slice $x_{A_{t+1}}$ at $t+1$. This seems to speak against the idea of starting from entities in the first place. If actions explicitly require a connection between them from one step to the next why have an additional notion of entity? Can we not only define actions (and perceptions) in this case and either have the entity emerge or ignored it completely? This is a valid approach and is sometimes discussed [3] but not much in keeping with the "entity-first" point of view in this thesis.

2. The notion of ι-entity is not strong enough to ensure that there are internal relations between all parts of the entities. This is possible but would be surprising. The requirement that all possible partitions have lower probability than the whole seems like a quite strong condition.

[3] Personal communication with Nathaniel Virgo.

3. The situation is not really a problem but is just not intuitive. We have also mentioned the time-step skipping ι-entities which are not intuitive in a Markov chain. It may be that the global way that ι-entities are defined connects the parts of entities in subtle ways. Maybe the action that seemed random above is not that random after all if seen from a global perspective (here "global" includes the entire Bayesian network, all time-steps and degrees of freedom). This ,however, still questions our motivation for our action definition. There we argued that if the environments are equal and the next time-slices different then the difference at the next time-step must originate from the agent (which it does not in the above example) or is random (which must then be wrong). So there either must be a third possibility due to some global effects or we have to accept that randomness is a kind of (proto-)action. A third possibility seems strange to us but on the other hand there are sometimes strange effects in probability and information theory.

4. Goal-directedness saves the situation. We expect that goal-directedness induces a connection between the entity perceptions and the entity actions. The entity actions have to somehow be adapted to the entity perceptions in order to get goal-directed behaviour. Then the example of an action above could stay an "action" or proto-action but the according entity could never be an agent since it cannot be goal-directed. This view still questions our motivation for the entity action definition. Again it seems we would have to accept random events as actions.

None of these possibilities is unproblematic. So more research is needed.

A third thing concerns non-interpenetration. We have seen that it provides uniquely defined branch-morphs and perceptions. However the ι-entities do interpenetrate. This leads to multiple coexisting but different perceptions for the same entity. There might be ways to define perception in a way similar to ours that deals with interpenetrating entities in a unique way and there might be completely different definitions of perception where the problem does not occur. However, the solution closest to the work in this thesis might be to get rid of non-interpenetrating entities. There could be well motivated ways to only use a (non-interpenetrating) subset of all ι-entities as an entity set. The partitions in the disintegration hierarchy all partition the system into non-interpenetrating blocks. We have in this thesis considered all blocks of all partitions together as the ι-entities. It is maybe not too far fetched to select blocks from only a few partitions or even a single partition as the entity-set. For example only blocks of partitions that are in some way on the same "scale" as others e.g. same cardinality, same block size, etc. This is related to the question of whether only entities on the same "scale" cannot interpenetrate but those on different "scales" can. We could also try to settle for a single partition, possibly the one in the refinement-free disintegration hierarchy with the absolute minimum of SLI. This would lead to non-interpenetration. This is future research.

There are also further requirements for entities that may be derived from phenomena associated to living organisms. Examples of such phenomena are

birth, death, growth, replication, etc. The implications of those will also be investigated in the future.

On the more technical side we would like to extend our definition of the branch-morph and try to generalise information theoretic measures for it.

Finally, there is the lack of a definition of *entity goal-directedness*. Even in the event that the definitions presented in this thesis turn out to be correct, entity goal-directedness has to weave the entity actions and entity perceptions together. So that they make some kind of sense as a sequence. This is essential if we want to give a full account of agents in lawful systems.

Chapter 7

Conclusion

The overarching goal of this thesis was to further improve our understanding of how the notion of agents can be compatible with a lawful universe. For this we attempted further formalisations of the notions identified in the literature as constitutive of agents. These are entities, perception, action, and goal-directedness. We made no contribution to the formalisation of goal-directedness.

With respect to entities we contributed a novel measure of integration called complete local integration (CLI). We have proposed this also as possible basis for a definition of entities. However, we first analysed this notion in the more general context of Bayesian networks. CLI is based on the specific local integration (SLI). We proved upper bounds constructively and constructed a candidate for a lower bound for SLI. We defined the disintegration-hierarchy and its refinement free version. Using these we revealed a relation between the finest partitions of global trajectories achieving a certain SLI value and the CLI of the blocks of these partitions. All blocks of such partitions are completely locally integrated. Conversely, all completely locally integrated spatiotemporal patterns are a block in at least one such partition. We expect that this result can be used to further investigate CLI and SLI formally. We presented an interpretation of this result from coding theory. This still needs rigorous analysis but seems promising (Section 5.3.5.3).

We also established the transformation of SLI under permutations of nodes in Bayesian networks and showed how they can be applied for the explanation of the disintegration hierarchy in an example Chapter 6. The transformation properties are also important for future theoretical work.

In Chapter 5 we presented three criteria for formal definitions of entities. These are compositionality, degree of freedom traversal, and counterfactual variation. Counterfactual variation has two realisations: in value and in extent. Based on these criteria we concluded that sets of random variables are not suitable for general agent definitions. We therefore defined general entity-sets as subsets of the set of all spatiotemporal patterns of a multivariate Markov chain. This definition turned out to be of considerable conceptual value. The main reason for this is that the perception-action loop also has an associated entity-set. By using this entity-set the new notions of entity action

and entity perception defined for arbitrary entity-sets naturally transform into notions known from the perception-action loop literature. The perception-action loop entity set is exhaustive, does not traverse degrees of freedom and only varies counterfactually in value. The general entity sets can vary counterfactually in extent, traverse degrees of freedom freely, and need not be exhaustive.

Our entity action definition was shown to imply non-heteronomy, an information theoretically defined notion related to autonomy due to Bertschinger et al. (2008). Entity perception was shown to correspond to a standard concept of perception in the perception-action loop. Entity perception is based on our construction of the branch-morph. This is a generalisation of the probability of the next agent state given current agent and environment states. This conditional probability distribution underlies the standard construction of perception in the perception-action loop. However, it also occurs in other concepts defined for the perception-action loop. The generalisation of such notions should now be feasible. Constructions similar to the branch-morph for other conditional probability distributions are also thinkable. We hope that this will lead to a generalised theory of perception and action for entity-sets.

During the construction of the branch-morph it became clear that it is uniquely defined if the entity-set satisfies non-interpenetration. This notion therefore seems worth investigating with respect to an axiom for entity sets. We were also able to show that the assumptions we need to make on the co-perception environments (which are the environments that can be perceived or distinguished by an entity) only appear stronger than the assumptions inherent in the perception-action loop and are actually equivalent. This is further support for our method of generalisation.

We noted that both entity actions and entity perception have multi-time-step analogues. Similar constructions are also used in perception-action loops. We also noted in the discussion of related work that our notion of perception can be seen as a formal generalisation to stochastic settings of the macroperturbations of the cognitive domain as defined for the glider in Beer (2014b).

Concerning simulation results we have calculated the disintegration hierarchies, visualised them and explained their structure using the SLI symmetry theorems. We also calculated the ι-entities (the completely locally integrated spatiotemporal patterns) for both example systems. For the example system with noise term we selected three trajectories. These results support our formal results.

With respect to the usefulness of the ι-entities as an entity-set the results are insufficient for a final assessment. As expected we found ι-entities that exhibit compositionality in time and space, counterfactual variation in value and extend, and degree of freedom traversal. However, we also found counter intuitive examples of ι-entities that skip a time-step completely. In other words they disappear for a time-step and reappear again. This could be due to the small system size but this is not certain and needs more investigation.

Concerning entity action and entity perception of the ι-entities. We found entity actions in value and extent. However we also find an example of an extent action that is counter intuitive and questions either our notion of entity

actions or the ι-entities. A decisive result is still lacking however.

Concerning perception, we found perceptions only where we expected it i.e. not for an independent process. However we could not define perceptions uniquely. A unique construction is guaranteed for non-interpenetrating entity-sets. However the ι-entities turned out to allow interpenetration. As expected we then also found non-exclusive co-perception entities. In order to still get perceptions we used the approach that only relies on a set ζ of mutually exclusive co-perception entities to define the branch-morph. We obtained perceptions but we also saw that another choice of a co-perception entity can lead to a (quantitatively) different perception. This further confirmed that without non-interpenetration perception is not necessarily uniquely defined.

Finally, we showed that the same ι-entity can perform entity actions and entity perception at the same time-step t.

All together the ι-entities need further investigation. One interesting next step is to use only the partition with the lowest SLI value for each trajectory as a source for entities. Such an entity-set would be non-interpenetrating. The other big question concerns the strange action we found. This also must be investigated further.

Bibliography

Albantakis, L., Hintze, A., Koch, C., Adami, C., and Tononi, G. (2014). Evolution of Integrated Causal Structures in Animats Exposed to Environments of Increasing Complexity. *PLoS Comput Biol*, 10(12):e1003966. 17

Amari, S.-I. (2001). Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711. 11, 114

Anthony, T., Polani, D., and Nehaniv, C. L. (2009). Impoverished Empowerment: 'Meaningful' Action Sequence Generation through Bandwidth Limitation. In Kampis, G., Karsai, I., and Szathmáry, E., editors, *Advances in Artificial Life. Darwin Meets von Neumann*, Lecture Notes in Computer Science, pages 294–301. Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-21314-4_37. 18

Ay, N. (2015). Information Geometry on Complexity and Stochastic Interaction. *Entropy*, 17(4):2432–2458. 12, 71, 72, 114

Ay, N., Bernigau, H., Der, R., and Prokopenko, M. (2012). Information-driven self-organization: the dynamical system approach to autonomous robot behavior. *Theory in Biosciences*, 131(3):161–179. 5, 18, 122

Ay, N. and Zahedi, K. (2014). On the Causal Structure of the Sensorimotor Loop. In Prokopenko, M., editor, *Guided Self-Organization: Inception*, number 9 in Emergence, Complexity and Computation, pages 261–294. Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-53734-9_9. 52

Balduzzi, D. (2011). Detecting emergent processes in cellular automata with excess information. *Advances in Artificial Life, ECAL*, abs/1105.0158. 3, 7, 12, 14, 15, 16, 54

Balduzzi, D. and Tononi, G. (2008). Integrated information in discrete dynamical systems: Motivation and theoretical framework. *PLoS Comput Biol*, 4(6):e1000091. 4, 12, 15, 71

Barandiaran, X. E., Paolo, E. D., and Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior*, 17(5):367–386. 2, 95, 96

Bartlett, S. and Bullock, S. (2015). Emergence of Competition between Different Dissipative Structures for the Same Free Energy Source. In *Proceedings*

of the European Conference on Artificial Life, pages 415–422. The MIT Press. 2, 5, 99, 101, 103

Bartlett, S. and Bullock, S. (2016). A Precarious Existence: Thermal Homeostasis of Simple Dissipative Structures. pages 608–615. The MIT Press. 2, 99, 101

Beer, R. D. (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence*, 72(1-2):173–215. 5, 18, 52, 53, 96, 122

Beer, R. D. (2014a). Characterizing autopoiesis in the game of life. *Artificial Life*, 21(1):1–19. 2, 3, 21, 22, 23, 25, 99

Beer, R. D. (2014b). The cognitive domain of a glider in the game of life. *Artificial Life*, 20(2):183–206. 2, 3, 5, 6, 7, 22, 23, 99, 104, 105, 181

Bertschinger, N., Olbrich, E., Ay, N., and Jost, J. (2006). Information and closure in systems theory. In *Explorations in the complexity of possible life : abstracting and synthesizing the principles of living systems - Proceedings of the 7th German Workshop on Artificial Life*, pages 9–19. Jena. 2, 5, 18, 52, 53, 56, 142, 147

Bertschinger, N., Olbrich, E., Ay, N., and Jost, J. (2008). Autonomy: An information theoretic perspective. *Biosystems*, 91(2):331–345. 2, 5, 6, 10, 18, 19, 20, 52, 53, 94, 122, 141, 142, 143, 144, 147, 181

Biehl, M., Ikegami, T., and Polani, D. (2016). Towards information based spatiotemporal patterns as a foundation for agent representation in dynamical systems. In *Proceedings of the Artificial Life Conference 2016*, pages 722–729. The MIT Press. 100

Biehl, M. and Polani, D. (2015). Apparent actions and apparent goal-directedness. In *Proceedings of the European Conference on Artificial Life 2015*, pages 511–511, York, UK. The MIT Press. 116

Bullen, P. S. (2003). *Handbook of Means and Their Inequalities*. Springer Science & Business Media. 67

Busic, A., Mairesse, J., and Marcovici, I. (2010). Probabilistic cellular automata, invariant measures, and perfect sampling. *arXiv:1010.3133 [cs, math]*. arXiv: 1010.3133. 35

Ceccherini-Silberstein, T. and Coornaert, M. (2009). Cellular Automata and Groups. In Ph. D, R. A. M., editor, *Encyclopedia of Complexity and Systems Science*, pages 778–791. Springer New York. DOI: 10.1007/978-0-387-30440-3_52. 35, 78

Chopard, B. (2009). Cellular Automata Modeling of Physical Systems. In Ph. D, R. A. M., editor, *Encyclopedia of Complexity and Systems Science*, pages 865–892. Springer New York. DOI: 10.1007/978-0-387-30440-3_57. 99

Christensen, W. D. and Hooker, C. A. (2000). Autonomy and the emergence of intelligence: Organised interactive construction. *Communication and Cognition-Artificial Intelligence*, 17(3-4):133–157. 96

Conway, J. (1970). The game of life. *Scientific American*, 223(4):4. 49, 99

Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory*. Wiley-Interscience, Hoboken, N.J. 113, 115

De Bruijn, N. G. (1970). *Asymptotic methods in analysis*, volume 4. Courier Corporation. 150

Der, R., Pasemann, F., and Steinmetz, U. (1999). Homeokinesis : a new principle to back up evolution with learning. In Mohammadian, M., editor, *Computational Intelligence for Modelling, Control and Automation*, volume 55 of *Concurrent Systems Engineering Series*, pages 43 – 47. IOS Press, Amsterdam [u.a.]. 122

England, J. L. (2013). Statistical physics of self-replication. *The Journal of Chemical Physics*, 139(12):121923. 99

Flecker, B., Alford, W., Beggs, J. M., Williams, P. L., and Beer, R. D. (2011). Partial information decomposition as a spatiotemporal filter. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 21(3):037104. 3, 12, 13

Franklin, S. and Graesser, A. (1997). Is it an agent, or just a program?: A taxonomy for autonomous agents. In *Intelligent Agents III Agent Theories, Architectures, and Languages*, pages 21–35. Springer. 96

Friston, K. (2013). Life as we know it. *Journal of The Royal Society Interface*, 10(86). 3, 23, 25, 26, 99

Froese, T., Virgo, N., and Ikegami, T. (2014). Motility at the origin of life: Its characterization and a model. *Artificial Life*, 20(1):55–76. 2, 5, 99, 101, 103

Froese, T. and Ziemke, T. (2009). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence*, 173(3–4):466–500. 96

Gallois, A. (2015). Identity Over Time. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Winter 2015 edition. 109

Gilmore, C. (2014). Location and mereology. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2014 edition. 109, 136

Grassberger, P. (1984). Chaos and diffusion in deterministic cellular automata. *Physica D: Nonlinear Phenomena*, 10(1):52–58. 12

Grätzer, G. (2011). *Lattice Theory: Foundation*. Springer, Basel ; New York, 2011 edition edition. 28, 29, 31, 32, 33

Hanson, J. E. and Crutchfield, J. P. (1992). The attractor—basin portrait of a cellular automaton. *Journal of Statistical Physics*, 66(5-6):1415–1462. 12

Hawley, K. (2015). Temporal parts. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2015 edition. 110

Hoel, E. P., Albantakis, L., and Tononi, G. (2013). Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences of the United States of America*, 110(49):19790–19795. 3, 4, 16, 20

Ikegami, T. and Taiji, M. (1998). Uncertainty, possible worlds and coupled dynamical recognizers. http://sacral.c.u-tokyo.ac.jp/pdf/ikegami_polytechnica_1998.pdf. 144

Kauffman, S. and Clayton, P. (2006). On emergence, agency, and organization. *Biology and Philosophy*, 21(4):501–521. 96

Kauffman, S. A. (2000). *Investigations*. Oxford University Press, Oxford; New York. 96

Klyubin, A., Polani, D., and Nehaniv, C. (2004). Organization of the information flow in the perception-action loop of evolved agents. In *2004 NASA/DoD Conference on Evolvable Hardware, 2004. Proceedings*, pages 177–180. 5, 52, 122

Klyubin, A., Polani, D., and Nehaniv, C. (2005). Empowerment: a universal agent-centric measure of control. In *The 2005 IEEE Congress on Evolutionary Computation, 2005*, volume 1, pages 128–135 Vol.1. 5, 18

Kolchinsky, A. and Rocha, L. M. (2011). Prediction and modularity in dynamical systems. *Advances in Artificial Life, ECAL*, pages 423–430. 115

Kondepudi, D., Kay, B., and Dixon, J. (2015). End-directed evolution and the emergence of energy-seeking behavior in a complex system. *Physical Review E*, 91(5):050902. 99

Krakauer, D., Bertschinger, N., Olbrich, E., Ay, N., and Flack, J. C. (2014). The information theory of individuality. *arXiv:1412.2447 [q-bio]*. arXiv: 1412.2447. 18, 19, 20, 103, 135

Langton, C. (1989). Artificial life. In Langton, C., editor, *Artificial Life*, pages 1–47. Addison-Wesley. 2

Legg, S. and Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *arXiv:0712.3329 [cs]*. arXiv: 0712.3329. 19

Lizier, J. T. (2012). *The Local Information Dynamics of Distributed Computation in Complex Systems*. Springer Science & Business Media. 11, 12, 14, 15, 114

Lizier, J. T., Prokopenko, M., and Zomaya, A. Y. (2008). Local information transfer as a spatiotemporal filter for complex systems. *Physical Review E*, 77(2):026110. 3, 12, 13

Lizier, J. T., Prokopenko, M., and Zomaya, A. Y. (2014). A framework for the local information dynamics of distributed computation in complex systems. In Prokopenko, M., editor, *Guided Self-Organization: Inception*, number 9 in Emergence, Complexity and Computation, pages 115–158. Springer Berlin Heidelberg. 122

Lucretius (2007). *The Nature of Things*. Penguin UK. 117

Lungarella, M., Pegors, T., Bulwinkle, D., and Sporns, O. (2005). Methods for quantifying the informational structure of sensory and motor data. *Neuroinformatics*, 3(3):243–262. 122

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press. 112, 113

Maes, P. (1993). Modeling adaptive autonomous agents. *Artificial Life*, 1(1-2):135–162. 96

Maturana, H. R. and Varela, F. J. (1980). *Autopoiesis and cognition: the realization of the living*. Springer. 2, 3, 21, 22

McGill, W. J. (1954). Multivariate information transmission. *Psychometrika*, 19(2):97–116. 11, 114

McGregor, S. (2016). A More Basic Version of Agency? As If! *Lecture Notes in Computer Science*, pages 183–194. 1

Ng, A. Y. and Russell, S. (2000). Algorithms for inverse reinforcement learning. In *in Proc. 17th International Conf. on Machine Learning*, pages 663–670. Morgan Kaufmann. 20

Nietzsche, F. W. (1892). *Zur Genealogie der Moral : eine Streitschrift*. Leipzig : C.G. Naumann. 117

Oizumi, M., Albantakis, L., and Tononi, G. (2014). From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Comput Biol*, 10(5):e1003588. 3, 4, 12, 15, 16

Orseau, L. and Ring, M. (2012). Space-Time Embedded Intelligence. In Bach, J., Goertzel, B., and Iklé, M., editors, *Artificial General Intelligence*, number 7716 in Lecture Notes in Computer Science, pages 209–218. Springer Berlin Heidelberg. 19, 20

Pantelis, P. C., Baker, C. L., Cholewiak, S. A., Sanik, K., Weinstein, A., Wu, C.-C., Tenenbaum, J. B., and Feldman, J. (2014). Inferring the intentional states of autonomous virtual agents. *Cognition*, 130(3):360–379. 20

Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge University Press. 42

Pemmaraju, S. and Skiena, S. (2009). *Computational Discrete Mathematics: Combinatorics and Graph Theory with Mathematica®*. Cambridge University Press. 154

Perunov, N., Marsland, R., and England, J. (2014). Statistical Physics of Adaptation. *arXiv:1412.1875 [cond-mat, physics:physics, q-bio]*. arXiv: 1412.1875. 99

Pivato, M. (2007). Defect particle kinematics in one-dimensional cellular automata. *Theoretical Computer Science*, 377(1–3):205–228. 12

Sagawa, T. (2012). *Thermodynamics of Information Processing in Small Systems*. Springer, New York, 2013 edition edition. 99

Saitta, A. M. and Saija, F. (2014). Miller experiments in atomistic computer simulations. *Proceedings of the National Academy of Sciences*, 111(38):13768–13773. 99

Salge, C., Glackin, C., and Polani, D. (2014). Changing the environment based on empowerment as intrinsic motivation. *Entropy*, 16(5):2789–2819. 5

Salge, C. and Polani, D. (2011). Digested information as an information theoretic motivation for social interaction. *Journal of Artificial Societies and Social Simulation*, 14(1):5. 18, 52

Schlosser, M. (2015). Agency. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Fall 2015 edition. 96, 119

Schmickl, T., Stefanec, M., and Crailsheim, K. (2016). How a life-like system emerges from a simplistic particle motion law. *Scientific Reports*, 6:37969. 5, 99, 101, 102, 103

Seth, A. K. (2010). Measuring autonomy and emergence via granger causality. *Artif. Life*, 16(2):179–196. 18, 122

Shalizi, C. R. (2001). *Causal architecture, complexity and self-organization in time series and cellular automata*. PhD thesis, University of Wisconsin–Madison. 7, 41, 54

Shalizi, C. R., Haslinger, R., Rouquier, J.-B., Klinkner, K. L., and Moore, C. (2006). Automatic filters for the detection of coherent structure in spatiotemporal systems. *Physical Review E*, 73(3):036104. 3, 12, 13, 14

Shalizi, C. R. and Moore, C. (2003). What is a macrostate? Subjective observations and objective dynamics. *arXiv preprint cond-mat/0303625*. 99

Smithers, T. (1995). Are autonomous agents information processing systems? In Steels, L. and Brooks, R. A., editors, *The Artificial Life Route to Artificial Intelligence: Building Embodied, Situated Agents*, pages 123–162. Lawrence Erlbaum Associates, Hillsdale, NJ, USA. 96

Still, S., Sivak, D. A., Bell, A. J., and Crooks, G. E. (2012). The thermodynamics of prediction. arXiv e-print 1203.3271. Phys. Rev. Lett. 109, 120604 (2012). 18, 99

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press. 18

Tishby, N. and Polani, D. (2011). Information Theory of Decisions and Actions. In Cutsuridis, V., Hussain, A., and Taylor, J. G., editors, *Perception-Action Cycle*, Springer Series in Cognitive and Neural Systems, pages 601–636. Springer New York. DOI: 10.1007/978-1-4419-1452-1_19. 18

Toffoli, T. (1984). Cellular automata as an alternative to (rather than an approximation of) differential equations in modeling physics. *Physica D: Nonlinear Phenomena*, 10(1):117 – 127. 2

Tononi, G. (2001). Information measures for conscious experience. *Archives italiennes de biologie*, 139(4):367. 4, 12, 15

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5:42. 4, 12, 15, 71, 114

Tononi, G. and Sporns, O. (2003). Measuring information integration. *BMC Neuroscience*, 4:31. 4, 15, 71

Tononi, G., Sporns, O., and Edelman, G. M. (1994). A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Sciences*, 91(11):5033–5037. 11, 12

Turing, A. M. (1952). The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 237(641):37–72. 99

Varela, F. J. (1979). *Principles of Biological Autonomy*. Monograph Collection (Matt - Pseudo). 21, 22

Virgo, N. (2011). *Thermodynamics and the Structure of Living Systems*. University of Sussex. Unpublished PhD thesis. 2, 5, 99, 101

von Eitzen, H. (2016). Prove $(1-(1-q)/n)^n \geqslant q$ for $0 < q < 1$ and $n \geqslant 2$ a natural number. Mathematics Stack Exchange. URL:http://math.stackexchange.com/q/1974262 (version: 2016-10-18). 67

Von Uexküll, J. (1920). *Theoretische biologie*. Paetel. 5, 52

Wilson, G. and Shpall, S. (2012). Action. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Summer 2012 edition. 1, 116, 117

Wolfram, S. (1983). Statistical mechanics of cellular automata. *Reviews of modern physics*, 55(3):601. 49

Wolfram, S. (1984). Computation theory of cellular automata. *Communications in mathematical physics*, 96(1):15–57. 12

Zahedi, K. and Ay, N. (2013). Quantifying morphological computation. *Entropy*, 15(5):1887–1915. arXiv:1301.6975 [cs, math]. 5, 18, 52, 147

Zahedi, K., Ay, N., and Der, R. (2010). Higher coordination with less control—a result of information maximization in the sensorimotor loop. *Adaptive Behavior*, 18(3-4):338–355. 18, 52

Zenil, H., Soler-Toscano, F., Delahaye, J.-P., and Gauvrit, N. (2015). Two-dimensional Kolmogorov complexity and an empirical validation of the Coding theorem method by compressibility. *PeerJ Computer Science*, 1:e23. 17