entropy

MDPI

Article

Specific and Complete Local Integration of Patterns in Bayesian Networks

Martin Biehl [1,2,*], Takashi Ikegami [3] and Daniel Polani [2]

[1] Araya Incorporation, 2F Mori 15 Building, 2-8-10 Toranomon, Minato-ku, Tokyo 105-0001, Japan
[2] School of Computer Science, University of Hertfordshire, Hatfield AL10 9AB, UK; d.polani@herts.ac.uk
[3] Department of General Systems Studies, University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan; ikeg@sacral.c.u-tokyo.ac.jp
* Correspondence: martin@araya.org; Tel.: +81-3-6550-9977

Academic Editor: Mikhail Prokopenko
Received: 19 March 2017; Accepted: 12 May 2017; Published: 18 May 2017

Abstract: We present a first formal analysis of specific and complete local integration. Complete local integration was previously proposed as a criterion for detecting entities or wholes in distributed dynamical systems. Such entities in turn were conceived to form the basis of a theory of emergence of agents within dynamical systems. Here, we give a more thorough account of the underlying formal measures. The main contribution is the disintegration theorem which reveals a special role of completely locally integrated patterns (what we call *ι-entities*) within the trajectories they occur in. Apart from proving this theorem we introduce the disintegration hierarchy and its refinement-free version as a way to structure the patterns in a trajectory. Furthermore, we construct the least upper bound and provide a candidate for the greatest lower bound of specific local integration. Finally, we calculate the ι-entities in small example systems as a first sanity check and find that ι-entities largely fulfil simple expectations.

Keywords: identity over time; Bayesian networks; multi-information; entity; persistence; integration; emergence; naturalising agency

1. Introduction

This paper investigates a formal measure and a corresponding criterion we developed in order to capture the notion of *wholes* or *entities* within Bayesian networks in general and multivariate Markov chains in particular. The main focus of this paper is to establish some formal properties of this criterion.

The main intuition behind wholes or entities is that combinations of some events/phenomena in space(-time) can be considered as more of a *single* or coherent "thing" than combinations of other events in space(-time). For example, the two halves of a soap bubble (The authors thank Eric Smith for pointing out the example of a soap bubble.) together seem to form more of a single thing than one half of a floating soap bubble together with a piece of rock on the ground. Similarly, the soap bubble at time t_1 and the "same" soap bubble at t_2 seem more like temporal parts of the *same* thing than the soap bubble at t_1 and the piece of rock at t_2. We are trying to formally define and quantify what it is that makes some spatially and temporally extended combinations of parts entities but not others.

We envisage spatiotemporal entities as a way to establish not only the problem of *spatial identity* but also that of *temporal identity* (also called *identity over time* [1]). In other words, in addition to determining which events in "space" (e.g., which values of different degrees of freedom) belong to the same structure spatiotemporal entities should allow the identification of the structure at a time t_2 that is the future (or past if $t_2 < t_1$) of a structure at time t_1. Given a notion of identity over time, it becomes possible to capture which things persist and in what way they persist. Without a notion of identity over

time, it seems persistence is not defined. The problem is how to decide whether something persisted from t_1 to t_2 if we cannot tell what at t_2 would count as the future of the original thing.

In everyday experience problems concerning identity over time are not of great concern. Humans routinely and unconsciously connect perceived events to spatially and temporally extended entities. Nonetheless, the problem has been known since ancient times, in particular with respect to artefacts that exchange their parts over time. A famous example is the Ship of Theseus which has all of its planks exchanged over time. This leads to the question whether it is still the *same* ship. From the point of view of physics and chemistry living organisms also exchange their parts (e.g., the constituting atoms or molecules) over time. In the long term we hope our theory can help to understand identity over time for these cases. For the moment, we are particularly interested in identity over time in formal settings like cellular automata, multivariate Markov chains, and more generally dynamical Bayesian networks. In these cases a formal notion of spatiotemporal entities (i.e., one defining spatial and temporal identity) would allow us to investigate persistence of entities/individuals formally. The persistence (and disappearance) of individuals are in turn fundamental to Darwinian evolution [2,3]. This suggests that spatiotemporal entities may be important for the understanding of the emergence of Darwinian evolution in dynamical systems.

Another area in which a formal solution to the problem of identity over time, and thereby entities (In the following, if not stated otherwise, we always mean *spatiotemporal* entities when we refer to entities.), might become important is a theory of intelligent agents that are space-time embedded as described by Orseau and Ring [4]. Agents are examples of entities fulfilling further properties e.g., exhibition of actions, and goal-directedness (cf. e.g., [5]). Using the formalism of reinforcement learning Legg and Hutter [6] proposes a definition of intelligence. Orseau and Ring [4] argue that this definition is insufficient. They dismiss the usual assumption that the environment of the reinforcement agent cannot overwrite the agent's memory (which in this case is seen as the memory/tape of a Turing machine). They conclude that in the most realistic case there only ever is one memory that the agent's (and the environment's) data is embedded in. They note that the difference between agent and environment then disappears. Furthermore, that the policy of the agent cannot be freely chosen anymore, only the initial condition. In order to measure intelligence according to Legg and Hutter [6] we must be able to define reward functions. This seems difficult without the capability to distinguish the agent according to some criterion. Towards the end of their publication Orseau and Ring [4] propose to define a "heart" pattern and use the duration of its existence as a reward. This seems a too specific approach to us since it basically defines identity over time (of the heart pattern) as invariance. In more general settings a pattern that maintains a more general criterion of identity over time would be desirable. Ideally, this criterion would also not need a specifically designed heart pattern. Another advantage would be that reward functions different from lifetime could be used if the agent were identifiable. An entity criterion in the sense of this paper would be a step in this direction.

1.1. Illustration

In order to introduce the contributions of this paper we illustrate the setting of our work further.

This illustration should only be taken as a motivation for what follows and not be confused with a result. The reason we don't use a concrete example is simply that we lack the necessary computational means (which are considerable as we will discuss in Section 5).

Let us assume we are given the entire time-evolution (what we will call a *trajectory*) of some known multivariate dynamical system or stochastic process. For example, a trajectory of a one-dimensional elementary cellular automaton showing a glider collision like Figure 1a (This is produced by the rule 62 elementary cellular automaton with time increasing from left to right. However, this does not matter here. For more about this system see e.g., Boccara et al. [7]).

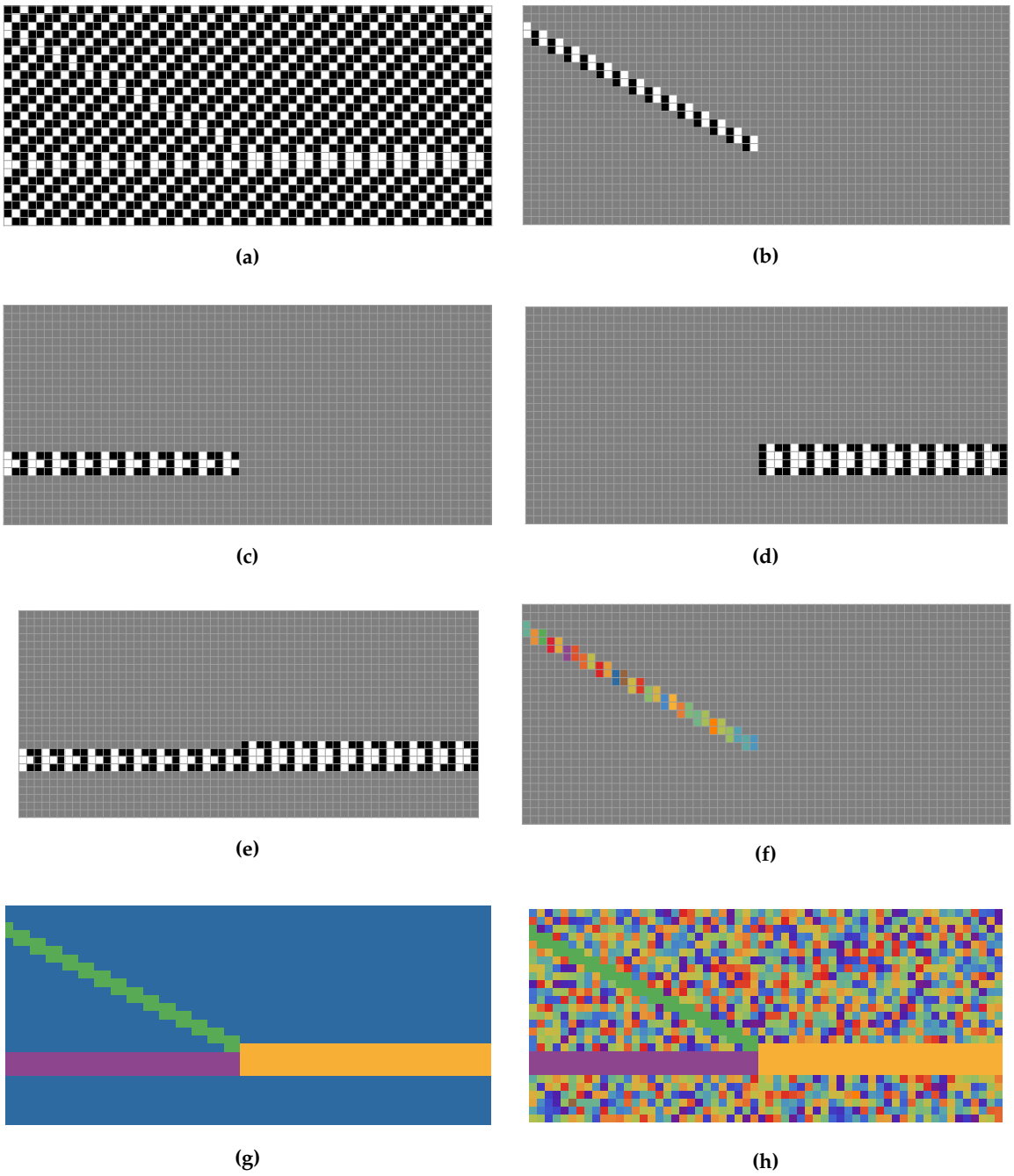


Figure 1. Illustration of concepts from this paper on the time-evolution (trajectory) of a one-dimensional elementary cellular automaton. Time-steps increase from left to right. None of the shown structures are derived from principles. They are manually constructed for illustrative purposes. In (**a**) we show the complete (finite) trajectory. Naively, two gliders can be seen to collide and give rise to a third glider; In (**b–d**) we show (spatiotemporal) patterns fixing the variables (allegedly) pertaining to a first, second, and a third glider; In (**e**) we show a pattern fixing the variables of what could be a glider that absorbs the first glider from before and maintains its identity; In (**f**) we show a partition into the time-slices of the pattern of the first glider; In (**g**) we show a partition of the trajectory with three parts coinciding with the gliders and one part encompassing the rest; In (**h**) we show again a partition with three parts coinciding with the gliders but now all other variables are considered as individual parts.

We take the point of view here argued for in previous work [8] that entities are phenomena that occur within trajectories and that they can be represented by (spatiotemporal) patterns. Patterns in

this sense fix a part of the variables in a trajectory to definite values and leave the rest undetermined. In Figure 1b–d we show such patterns that occur in Figure 1a with the undetermined variables coloured grey and the determined ones taking on those of the trajectory. Visually speaking, a pattern is a snippet from a trajectory that it occurs in.

From Figure 1a we would probably expect that what we are seeing are two gliders colliding and forming a third. However, it may also be that one of the gliders absorbs the other, maintains some form of identity, and only changes its appearance (e.g., it "grows"). This highlights the problem of identity over time. While the spatial identity of such patterns has been treated multiple times in the literature their identity of over time is rarely dealt with.

Our approach evaluates the "integration" of spatiotemporally extended patterns at once. According to our proposed entity-criterion a pattern is an ι-entity if, due to the dynamics of the system, every part of this pattern (which is again a pattern) makes all other parts more probable. Identity over time is then included since future parts have to make past parts more probable and vice versa. In principle this would allow us to detect if one of the gliders absorbs another one without loosing its identity. For example, this could result in an entity as in Figure 1e.

In order to detect entities the straightforward approach is to evaluate the entity-criterion for every spatiotemporal pattern in a given trajectory. Evaluating our entity-criterion of positive complete local integration (CLI) for a given pattern corresponds to splitting the pattern into parts in every possible way and calculating whether all the resulting parts make each other more probable. This means evaluating the specific local integration (SLI) with respect to all *partitions* of the set of variables occupied by the pattern.

1.2. Contributions

This paper contains four contributions.

We first give a more formal definition of patterns. Since each pattern uniquely specifies a set of trajectories (those trajectories that the pattern occurs in) one might be tempted to reduce the analysis to that of sets of trajectories. We show that this is not possible since not all sets of trajectories have a pattern that specifies them.

Second, we try to get a general intuition for the patterns whose parts make all other parts more probable. For this we show how to construct patterns that, for given probability of the whole pattern, achieve the least upper bound of specific local integration (SLI). These turn out to be patterns for which each part only occurs if and only if the whole pattern occurs. We also construct a pattern that, again for given probability of the whole pattern, has *negative* SLI. These pattern (which may achieve the greatest lower bound of SLI) occur if either the whole pattern occurs or the pattern occurs up to exactly one part of it, which does not occur.

Third, we prove the disintegration theorem. This is the main contribution. We saw that patterns are snippets of trajectories. We can also look at the whole trajectory as a single pattern. Like all patterns the trajectory can be split up into parts, i.e., partitioned, resulting in a set of patterns. Among the partitions we find examples such as those in Figure 1g,h. These are very particular partitions picking out the gliders among all possible parts. This suggests that finding such special partitions provides a (possibly different) notion of entities.

One intuition we might have is that entities are the most "independent" parts of a trajectory. In other words we could look for the partition whose parts make the other parts *less* probable. The disintegration theorem then shows that this approach again leads to the ι-entities. This shows that ι-entities do not only have an intuitive motivation but also play a particular role in the structure of probabilities of entire trajectories.

It is not directly the parts of the partitions that minimise SLI for a trajectory which are ι-entities. To get ι-entities we first classify all partitions of the trajectory according to their SLI value. Then within each such class we choose the partitions for which no refining partition (A refining partition is one that further partitions any of the parts of the original partition.) achieves an even lower level of SLI.

So according to the disintegration theorem a ι-entity is not only a pattern that is integrated with respect to every possible partition of the pattern but also a pattern that occurs in partitions that minimise (in a certain sense) the integration of trajectories.

A side effect of the disintegration theorem is that we naturally get a kind of hierarchy of ι-entities called the disintegration hierarchy. For each trajectory and its different levels of SLI we find different decompositions of the trajectory into ι-entities.

Fourth, we calculate the ι-entities and disintegration hierarchy for two simple example systems. Our example systems show that in general the partitions at a particular disintegration level are not unique. This means that there are overlapping ι-entities at those levels. Furthermore, the same ι-entity can occur on multiple levels of the disintegration.

We do not thoroughly discuss the disintegration hierarchies in this paper and postpone this to future publications. Here we only note that many entities in the real world occur within hierarchies as well. For example, animals are entities that are composed of cells which are themselves entities.

1.3. Related Work

We now give a quick overview of related work. More in depth discussions will be provided after we formally introduce our definitions.

To our knowledge the measure of CLI has been proposed for the first time by us in [8]. However, this publication contained none of the formal or numerical results in the present paper. From a formal perspective the measures of SLI and CLI are a combination of existing concepts. SLI localises multi-information [9,10] in the way proposed by Lizier [11] for other information theoretic measures. In order to get the CLI we apply the weakest-link approach proposed by Tononi and Sporns [12], Balduzzi and Tononi [13] to SLI.

Conceptually, our work is most closely related to Beer [14]. The notion of spatiotemporal patterns used there to capture blocks, blinkers, and gliders is equivalent to the *patterns* we define more formally here. This work also contains an informal entity-criterion that directly deals with identity over time (not only space). It differs significantly from our proposal as it depends on the re-occurrence of certain transitions at later times in a pattern whereas our criterion only depends on the probabilities of parts of the patterns without the need for any re-occurrences.

The *organisations* of chemical organisation theory [15] may also be interpreted as entity-criteria. In Fontana and Buss [15] these are defined in the following way:

> The observer will conclude that the system is an organisation to the extent that there is a compressed description of its objects and of their relations.

The direct intuition is different from ours and it is not clear to us in how far our entity-criterion is equivalent to this. This will be further investigated in the future.

It is worth noting that viewing entities/objects/individuals as patterns occurring within a trajectory is in contrast to an approach that models them as sets of random variables/stochastic processes (e.g., a set of cells in a CA in contrast to a set of specific values of a set of cells). An example of the latter approach are the information theoretic individuals of Krakauer et al. [16]. These individuals are identified using an information theoretic notion of autonomy due to Bertschinger et al. [17]. The latter notion of autonomy is also somewhat related to the idea of integration here. Autonomy contains a term that measures the degree to which a random variable representing an individual at timestep t determines the random variable representing it at $t+1$. Similarly, CLI requires that every part of an entity pattern makes every other part more probable, in the extreme case this means that every part determines that every other part of the pattern also occurs. However, formally autonomy evaluates random variables and not patterns directly.

At the most basic level the intuition behind entities is that some spatiotemporal patterns are more special than others. 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. [18] have argued most of the earlier definitions and methods [19–22] require previous knowledge about the patterns being looked for. They are therefore not suitable for a general definition of entities. More recent definitions based on information theory [18,23,24] do not have this limitation anymore. The difference to our entity-criterion is that they do not treat identity over time. They are well suited to identify gliders at each time-step for example, but if two gliders collide and give rise to a third glider as in Figure 1a these methods (by design) say nothing about the identity of the third glider. i.e., they cannot make a difference between a glider absorbing another one and two gliders producing a new one. While we have not been able to show that our approach actually makes such distinctions for gliders, it could do so in principle.

We note here that the approach of identifying individuals by Friston [25] using Markov blankets has the same shortcoming as the spatiotemporal filters. For each individual time-step it returns a partition of all degrees of freedom into internal, sensory, active, and external degrees. However, it does not provide a way to resolve ambiguities in the case of multiple such partitions colliding.

Among research related to integrated information theory (IIT) there are approaches (a first one by Balduzzi [26] and a more recently by Hoel et al. [27]) that can be used to determine specific spatiotemporal patterns in a trajectory. They can therefore be interpreted to define a notion of entities even if that is not their main goal. These approaches are aimed at establishing the optimal spatiotemporal coarse-graining to describe the dynamics of a system. For a given trajectory we can then identify the patterns that instantiate a macro-state/coarse-grain that is optimal according to their criterion.

In contrast to our approach the spatiotemporal grains are determined by their interactions with other grains. In our case the entities are determined first and foremost by their internal relations.

The consequence seems to be that a pattern can be an entity in one trajectory and not an entity in another even if it occurs in both. In our conception a pattern is an entity in all trajectories it occurs in.

2. Notation and Background

In this section we briefly introduce our notation for sets of random variables (Since every set of jointly distributed random variables can be seen as a Bayesian network and vice versa we use these terms interchangeably.) and their partition lattices.

In general, we use the convention that upper-case letters X, Y, Z are random variables, lower-case letters x, y, z are specific values/outcomes of random variables, and calligraphic letters $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ are state spaces that random variables take values in. Furthermore:

Definition 1. *Let $\{X_i\}_{i\in V}$ be a set of random variables with totally ordered finite index set V and state spaces $\{\mathcal{X}_i\}_{i\in V}$ respectively. Then for $A, B \subseteq V$ define:*

1. *$X_A := (X_i)_{i\in A}$ as the joint random variable composed of the random variables indexed by A, where A is ordered according to the total order of V,*
2. *$\mathcal{X}_A := \prod_{i\in A} \mathcal{X}_i$ as the state space of X_A,*
3. *$x_A := (x_i)_{i\in A} \in \mathcal{X}_A$ as a value of X_A,*
4. *$p_A : \mathcal{X}_A \to [0,1]$ as the probability distribution (or more precisely probability mass function) of X_A which is the joint probability distribution over the random variables indexed by A. If $A = \{i\}$ i.e., a singleton set, we drop the parentheses and just write $p_A = p_i$,*
5. *$p_{A,B} : \mathcal{X}_A \times \mathcal{X}_B \to [0,1]$ as the probability distribution over $\mathcal{X}_A \times \mathcal{X}_B$. Note that in general for arbitrary $A, B \subseteq V$, $x_A \in \mathcal{X}_A$, and $y_B \in \mathcal{X}_B$ this can be rewritten as a distribution over the intersection of A and B and the respective complements. The variables in the intersection have to coincide:*

$$p_{A,B}(x_A, y_B) := p_{A\setminus B, A\cap B, B\setminus A, A\cap B}(x_{A\setminus B}, x_{A\cap B}, y_{B\setminus A}, y_{A\cap B}) \tag{1}$$

$$= \delta_{x_{A\cap B}}(y_{A\cap B})\; p_{A\setminus B, A\cap B, B\setminus A}(x_{A\setminus B}, x_{A\cap B}, y_{B\setminus A}). \tag{2}$$

Here δ is the Kronecker delta (see Appendix A). If $A \cap B = \emptyset$ and $C = A \cup B$ we also write $p_C(x_A, y_B)$ to keep expressions shorter.

6. *$p_{B|A} : \mathcal{X}_A \times \mathcal{X}_B \to [0,1]$ with $(x_A, x_B) \mapsto p_{B|A}(x_B|x_A)$ as the conditional probability distribution over X_B given X_A:*

$$p_{B|A}(y_B|x_A) := \frac{p_{A,B}(x_A, y_B)}{p_A(x_A)}. \tag{3}$$

We also just write $p_B(x_B|x_A)$ if it is clear from context what variables we are conditioning on.

If we are given p_V we can obtain every p_A through marginalisation. In the notation of Definition 1 this is formally written:

$$p_A(x_A) = \sum_{\bar{x}_{V\setminus A} \in \mathcal{X}_{V\setminus A}} p_{A,V\setminus A}(x_A, \bar{x}_{V\setminus A}) \tag{4}$$

$$= \sum_{\bar{x}_{V\setminus A} \in \mathcal{X}_{V\setminus A}} p_V(x_A, \bar{x}_{V\setminus A}). \tag{5}$$

Next we define the partition lattice of a set of random variables. Partition lattices occur as a structure of the set of possible ways to split an object/pattern into parts. Subsets of the partition lattices play an important role in the disintegration theorem.

Definition 2 (Partition lattice of a set of random variables)**.** *Let $\{X_i\}_{i\in V}$ be a set of random variables.*

1. *Then its partition lattice $\mathfrak{L}(V)$ is the set of partitions of V partially ordered by refinement (see also Appendix B).*
2. *For two partitions $\pi, \rho \in \mathfrak{L}(V)$ we write $\pi \lhd \rho$ if π refines ρ and $\pi \lhd: \rho$ if π covers ρ. The latter means that $\pi \neq \rho$, $\pi \lhd \rho$, and there is no $\xi \in \mathfrak{L}(V)$ with $\pi \neq \xi \neq \rho$ such that $\pi \lhd \xi \lhd \rho$.*
3. *We write $\mathbf{0}$ for the zero element of a partially ordered set (including lattices) and $\mathbf{1}$ for the unit element.*
4. *Given a partition $\pi \in \mathfrak{L}(V)$ and a subset $A \subseteq V$ we define the restricted partition $\pi|_A$ of π to A via:*

$$\pi|_A := \{b \cap A : b \in \pi\}. \tag{6}$$

For some examples of partition lattices see Appendix B and for more background see e.g., Grätzer [28]. For our purpose it is important to note that the partitions of sets of random variables or Bayesian networks we are investigating are partitions of the index set V of these and not partitions of their state spaces $\mathcal{X}_V$.

3. Patterns, Entities, Specific, and Complete Local Integration

This section contains the formal part of this contribution.

First we introduce *patterns*. Patterns are the main structures of interest in this publication. Entities are seen as special kinds of patterns. The measures of specific local integration and complete local integration, which we use in our criterion for ι-entities, quantify notions of "oneness" of patterns. We give a brief motivation and show that while each pattern defines a set of "trajectories" of a set of random variables not every such set is defined by a pattern. This justifies studying patterns for their own sake.

Then we motivate briefly the use of specific and complete local integration (SLI and CLI) for an entity criterion on patterns. We then turn to more formal aspects of SLI and CLI. We first prove an upper bound for SLI and construct a candidate for a lower bound. We then go on to define the disintegration hierarchy and its refinement-free version. These structures are used to prove the main result, the *disintegration theorem*. This relates the SLI of whole trajectories of a Bayesian network to the CLI of parts of these trajectories and vice versa.

3.1. Patterns

This section introduces the notion of patterns. These form the basic candidate structures for entities.

The structures we are trying to capture by entities should be analogous to spatially and temporally extended objects we encounter in everyday life (e.g., soap bubbles, living organisms). These objects seem to occur in the single history of the universe that also contains us. The purpose of patterns is then to capture arbitrary structures that occur within single trajectories or histories of a multivariate discrete dynamical system (see Figure 2 for an example of a Bayesian network of such a system, any cellular automaton is also such a system).

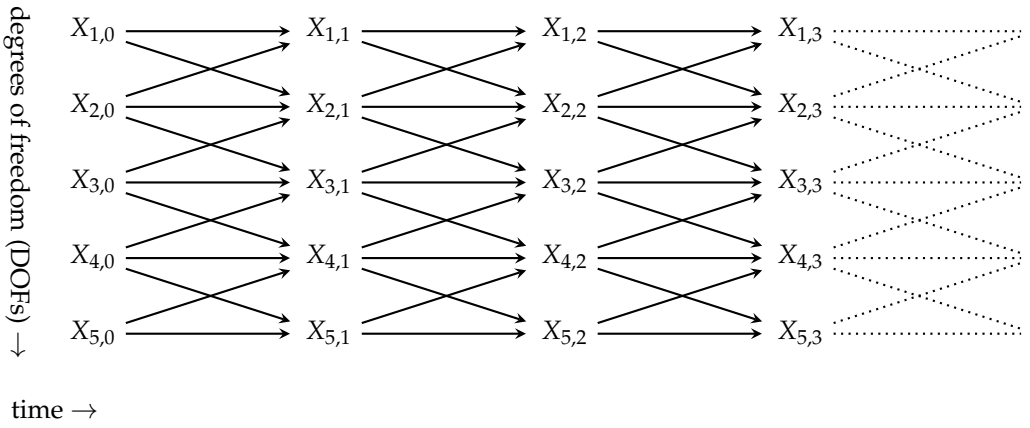
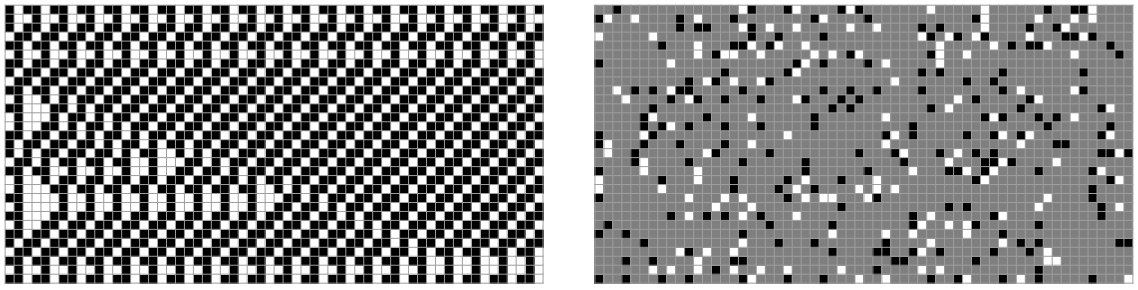


Figure 2. First time steps of a Bayesian network representing a multivariate dynamical system (or multivariate Markov chain) $\{X_i\}_{i \in V}$. Here we used $V = J \times T$ with J indicating spatial degrees of freedom and T the temporal extension. Then each node is indexed by a tuple (j, t) as shown. The shown edges are just an example, edges are allowed to point from any node to another one within the same or in the subsequent column.

We emphasise the single trajectory since many structures of interest (e.g., gliders) occur in some trajectories in some "places", in other trajectories in other "places" (compare e.g., Figures 1a and 3a), and in some trajectories not at all. We explicitly want to be able to capture such trajectory dependent structures and therefore choose patterns. Examples of formal structures for which it makes no sense to say that they occur within a trajectory are for example the random variables in a Bayesian network and, as we will see, general sets of trajectories of the Bayesian network.

Unlike entities, which we conceive of as special patterns that fulfil further criteria, patterns are formed by *any* combination of events at arbitrary times and positions. As an example, we might think of cellular automaton again. The time evolutions over multiple steps of the cells attributed to a glider see [14] for a principled way to attribute cells to theseas in Figure 1b,e should be patterns but also arbitrary choices of events in a trajectory as in Figure 3b.

In the more general context of (finite) Bayesian networks there may be no interpretation of time or space. Nonetheless, we can define that a trajectory in this case fixes every random variable to a particular value. We then define patterns formally in the following way.



(a) **(b)**

Figure 3. In (**a**) we show a trajectory of the same cellular automaton as in Figure 1 with a randomly chosen initial condition. The set of gliders and their paths occurring in this trajectory is clearly different from those in Figure 1a. In (**b**) we show an example of a random pattern that occurs in the trajectory of (**a**) and is probably not an entity in any sense.

Definition 3 (Patterns and trajectories)**.** *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.*

1. *A pattern at $A \subseteq V$ is an assignment*

$$X_A = x_A \tag{7}$$

   where $x_A \in \mathcal{X}_A$. If there is no danger of confusion we also just write x_A for the pattern $X_A = x_A$ at A.
2. *The elements x_V of the joint state space $\mathcal{X}_V$ are isomorphic to the patterns $X_V = x_V$ at V which fix the complete set $\{X_i\}_{i\in V}$ of random variables. Since they will be used repeatedly we refer to them as the* trajectories *of $\{X_i\}_{i\in V}$.*
3. *A pattern x_A is said to occur in trajectory $\bar{x}_V \in \mathcal{X}_V$ if $\bar{x}_A = x_A$.*
4. *Each pattern x_A uniquely defines (or captures) a set of trajectories $\mathcal{T}(x_A)$ via*

$$\mathcal{T}(x_A) = \{\bar{x}_V \in \mathcal{X}_V : \bar{x}_A = x_A\}, \tag{8}$$

   i.e., the set of trajectories that x_A occurs in.
5. *It is convenient to allow the empty pattern $x_\emptyset$ for which we define $\mathcal{T}(x_\emptyset) = \mathcal{X}_V$.*

Remarks:

- Note that for every $x_A \in \mathcal{X}_A$ we can form a pattern $\mathcal{X}_A = x_A$ so the set of all patterns is $\bigcup_{A\subseteq V} \mathcal{X}_A$.
- Our notion of patterns is similar to "patterns" as defined in [29] and to "cylinders" as defined in [30]. More precisely, these other definitions concern (probabilistic) cellular automata where all random variables have identical state spaces $\mathcal{X}_i = \mathcal{X}_j$ for all $i, j \in V$. They also restrict the extent of the patterns or cylinders to a single time-step. Under these conditions our patterns are isomorphic to these other definitions. However, we drop both the identical state space assumption and the restriction to single time-steps.

  Our definition is inspired by the usage of the term "spatiotemporal pattern" in [14,31,32]. There is no formal definition of this notion given in these publications but we believe that our definition is a straightforward formalisation. Note that these publications only treat the Game of Life cellular automaton. The assumption of identical state space is therefore implicitly made. At the same time the restriction to single time-steps is explicitly dropped.

Since every pattern defines a subset of $\mathcal{X}_V$, one could think that every subset of $\mathcal{X}_V$ is also a pattern. In that case studying patterns in a set of random variables $\{X_i\}_{i\in V}$ would be the same as studying subsets of its set of trajectories $\mathcal{X}_V$. However, the set of subsets of $\mathcal{X}_V$ defined by patterns

and the set of all subsets $2^{\mathcal{X}_V}$ (i.e., the power set) of $\mathcal{X}_V$ of a set of random variables $\{X_i\}_{i\in V}$ are not identical. Formally:

$$\bigcup_{B\subseteq V}\{\mathcal{T}(x_B)\subseteq\mathcal{X}_V : x_B\in\mathcal{X}_B\}\subseteq 2^{\mathcal{X}_V}. \tag{9}$$

While patterns define subsets of $\mathcal{X}_V$, not every subset of $\mathcal{X}_V$ is captured by a pattern. The difference of the two sets is characterised in Theorem 1 below. We first present a simple example of a subset $\mathcal{D}\in 2^{\mathcal{X}_V}$ that cannot be captured by a pattern.

Let $V=\{1,2\}$ 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 $\bar{x}_A=(1,1)$. Then let

$$\mathcal{D}:=\{x_A\cup\bar{x}_A\}=\{(0,0),(1,1)\}. \tag{10}$$

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

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

(11)

and verify that $\mathcal{D}$ is not among them. Before we formally characterise the difference, we define some extra terminology.

Definition 4. *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. For a subset $\mathcal{D}\subseteq\mathcal{X}_V$ the set $\mathcal{D}_A$ of all patterns at A that occur in one of the trajectories in $\mathcal{D}$ is defined as*

$$\mathcal{D}_A:=\{x_A\in\mathcal{X}_A : \exists\bar{x}_V\in\mathcal{D},\bar{x}_A=x_A\}. \tag{12}$$

So in the previous example $\mathcal{D}_{\{1\}}=\{0,1\}$, $\mathcal{D}_{\{2\}}=\{0,1\}$, $\mathcal{D}_{\{1,2\}}=\{(0,0),(1,1)\}$. In then get the following theorem which establishes the difference between the subsets of $\mathcal{X}_V$ captured by patterns and general subsets.

Theorem 1. *Given a set of random variables $\{X_i\}_{i\in V}$, a subset $\mathcal{D}\subseteq\mathcal{X}_V$ cannot be represented by a pattern of $\{X_i\}_{i\in V}$ if and only if there exists $A\subseteq V$ with $\mathcal{D}_A\subset\mathcal{X}_A$ (proper subset) and $|\mathcal{D}_A|>1$, i.e., if neither all patterns at A are possible nor a unique pattern at A is specified by $\mathcal{D}$.*

Proof. See Appendix D. □

We saw that in the previous example the subset $\mathcal{D}$ cannot be captured by a pattern. For $A=\{1\}$ we have $\mathcal{D}_{\{1\}}=\{0,1\}=\mathcal{X}_{\{1\}}$ and for $A=\{2\}$ we have $\mathcal{D}_{\{2\}}=\{0,1\}=\mathcal{X}_{\{2\}}$ so these do not fulfil the conditions of Theorem 1. However, for $A=\{1,2\}$ we have $\mathcal{D}_{\{1,2\}}=\{(0,0),(1,1)\}\subset\mathcal{X}_{\{1,2\}}$ and $|\mathcal{D}_{\{1,2\}}|>1$ so the conditions of Theorem 1 are fulfilled and as expected $\mathcal{D}$ cannot be captured by a pattern.

The proof of the following corollary shows how to construct a subset that cannot be represented by a pattern for all sets of random variables $\{X_i\}_{i\in V}$ with $|\mathcal{X}_V| > 2$.

Corollary 1. *Given a set of random variables* $\{X_i\}_{i\in V}$*, if* $|\mathcal{X}_V| > 2$ *then*

$$\bigcup_{B\in V} \{\mathcal{T}(x_B) \subseteq \mathcal{X}_V : x_B \in \mathcal{X}_B\} \subset 2^{\mathcal{X}_V} \tag{13}$$

(proper subset).

Proof. Choose $\mathcal{D} = \{x_V, y_V\} \in 2^{\mathcal{X}_V}$ with $y_V \in \{\bar{x}_V \in \mathcal{X}_V : \forall i \in V, \bar{x}_i \neq x_i\}$. Then for all $A \subseteq V$ we have $|\mathcal{D}_A| = 2$ and $\mathcal{D}_A \subset \mathcal{X}_A$. So $\mathcal{D}$ cannot be represented by a pattern according to Theorem 1 and so $\mathcal{D} \notin \bigcup_{B\in V}\{\mathcal{T}(x_B) \subseteq \mathcal{X}_V : x_B \in \mathcal{X}_B\}$. □

This means that in every set of random variables that not only consists of a single binary random variable there are subsets of $\mathcal{X}_V$ that cannot be captured by a pattern. We can interpret this result in the following way. Patterns were constructed to be structures that occur within trajectories. It then turned out that each pattern also defines a subset of all trajectories of a system. So for sets of trajectories captured by patterns it could make sense to say they "occur" within one trajectory. However, there are sets of trajectories that are not captured by patterns. For these sets of trajectories it would then not be well-defined to say that they occur within a trajectory. This is the reason we choose to investigate patterns specifically and not sets of trajectories.

3.2. Motivation of Complete Local Integration as an Entity Criterion

We proposed to use patterns as the candidate structures for entities since patterns comprise arbitrary structures that occur within single trajectories of multivariate systems. Here we heuristically motivate our choice of using positive complete local integration as a criterion to select entities among patterns. In general such a criterion would give us, for any Bayesian network $\{X_i\}_{i\in V}$ a subset $\mathfrak{E}(\{X_i\}_{i\in V}) \subseteq \bigcup_{A\subseteq V} \mathcal{X}_A$ of the patterns.

So what is an entity? We can rephrase the problem of finding an entity criterion by saying an entity is composed of parts that share the same identity. So if we can define when parts share the same identity we also define entities by finding all parts that share identity with some given part. For the moment, let us decompose (as is often done [33]) the problem of identity into two parts:

1. spatial identity and
2. temporal identity.

Our solution will make no distinction between these two aspects in the end. We note here that conceiving of entities (or objects) as composite of spatial and temporal parts as we do in this paper is referred to as four-dimensionalism or perdurantism in philosophical discussions (see e.g., [34]). The opposing view holds that entities are spatial only and endure 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. Parts of living organisms rarely exist without the rest of the living organisms also existing. For example, it is rare that an arm exists without a corresponding rest of a human body existing compared to an arm and the rest of a human body existing. The body (without arm) seems to make the existence of the arm more probable and vice versa. Similar relations between parts seem 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. 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? Does the existence of an arm make things more probable that are not parts of the corresponding body? For example, does the arm make the existence of some piece of rock more probable? Maybe to a small degree as without the existence of any rocks in the universe humans are probably impossible. However, this effect is much smaller than the increase of probability of the existence of the rest of the body due to the arm.

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. An arm at some instance of time does not make the existence of a rock at a subsequent instance much more probable. It does make the existence of a human body at a subsequent instance much more probable. So the human body at the second instance seems to be more like a future part of the arm than the rock. Switching now to patterns in sets of random variables we can easily formalise such intuitions. We required that for an entity every part of the structure, which is now a pattern x_O, makes every other part more probable. A part of a pattern is a pattern x_b with $b \subset O$. If we require that every part of a pattern makes every other part more probable then we can write that x_O is an entity if:

$$\min_{b \subset O} \frac{p_{O \setminus b}(x_{O \setminus b} | x_b)}{p_{O \setminus b}(x_{O \setminus b})} > 1. \tag{14}$$

This is equivalent to

$$\min_{b \subset O} \frac{p_O(x_O)}{p_{O \setminus b}(x_{O \setminus b}) p_b(x_b)} > 1. \tag{15}$$

If we write $\mathfrak{L}_2(O)$ for the set of all bipartitions of O we can rewrite this further as

$$\min_{\pi \in \mathfrak{L}_2(O)} \frac{p_O(x_O)}{\prod_{b \in \pi} p_b(x_b)} > 1. \tag{16}$$

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 pattern $x_O = (x_{b_1}, x_{b_2})$ 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

$$p_O(x_O) =: p_{b_1, b_2}(x_{b_1}, x_{b_2}) = p_{b_1}(x_{b_1}) p_{b_2}(x_{b_2}). \tag{17}$$

Which would give us

$$\frac{p_O(x_O)}{\prod_{b \in \pi} p_b(x_b)} = 1 \tag{18}$$

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

$$p_O(x_O) = \prod_{c \in \xi} p_c(x_c) \tag{19}$$

seems to suggest that the pattern x_O is not an entity but instead composite of three parts. We can therefore generalise Equation (16) to include all partitions $\mathfrak{L}(O)$ (see Definition 2) of O except the unit partition $\mathbf{1}_O$. Then we would say that x_O is an entity if

$$\min_{\pi \in \mathfrak{L}(O) \setminus \mathbf{1}_O} \frac{p_O(x_O)}{\prod_{b \in \pi} p_b(x_b)} > 1. \tag{20}$$

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. We then arrive at the following entity-criterion

$$\min_{\pi \in \mathfrak{L}(O) \setminus \mathbf{1}_O} \log \frac{p_O(x_O)}{\prod_{b \in \pi} p_b(x_b)} > 0. \tag{21}$$

where the left hand side is the complete local integration (CLI), the function minimised is the specific local integration (SLI), and the inequality provides the criterion for ι-entities. For reference, we define these notions formally. We begin with SLI which quantifies for a given partition π of a pattern in how far the probability of the whole pattern is bigger than its probability would be if the blocks of the partition would be independent.

Definition 5 (Specific local integration (SLI))**.** *Given a Bayesian network $\{X\}_{i \in V}$ and a pattern 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*

$$\mathrm{mi}_\pi(x_O) := \log \frac{p_O(x_O)}{\prod_{b \in \pi} p_b(x_b)}. \tag{22}$$

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

Definition 6 ((Complete) local integration)**.** *Given a Bayesian network $\{X_i\}_{i \in V}$ and a pattern x_O of this network the complete local integration $\iota(x_O)$ of x_O is the minimum SLI over the non-unit partitions $\pi \in \mathfrak{L}(O) \setminus \mathbf{1}_O$:*

$$\iota(x_O) := \min_{\pi \in \mathfrak{L}(O) \setminus \mathbf{1}_O} \mathrm{mi}_\pi(x_O). \tag{23}$$

We call a pattern 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 2) is that with respect to it every pattern has $\mathrm{mi}_{\mathbf{1}_O}(x_O) = 0$.
- Looking for a partition that minimises a measure of integration is known as the *weakest link approach* [35] to dealing with multiple partitions. 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 the expected value of SLI see Ay [35] and references therein. For our interpretation taking the average seems less well suited since requiring a positive average will allow SLI to be negative with respect to some partitions.

Definition 7 (ι-entity)**.** *Given a multivariate Markov chain $\{X_i\}_{i \in V}$ a pattern x_O is a ι-entity if*

$$\iota(x_O) > 0. \tag{24}$$

The entire set of ι-entities $\mathfrak{E}_\iota(\{X_i\}_{i \in V})$ is then defined as follows.

Definition 8 (*ci*-entity-set)**.** *Given a multivariate Markov chain $\{X_i\}_{i \in V}$ the ι-entity-set is the entity-set*

$$\mathfrak{E}_\iota(\{X_i\}_{i \in V}) := \{x_O \in \bigcup_{A \subseteq V} \mathcal{X}_A : \iota(x_O) > 0\}. \tag{25}$$

Next, we look at some interpretations that the introduction of the logarithm allows.

- A first consequence of introducing the logarithm is that we can now formulate the condition of Equation (24) 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 Equation (22) as:

$$\mathrm{mi}_\pi(x_O) = \log p_O(x_O) - \sum_{b\in\pi} \log p_b(x_b). \tag{26}$$

- Another side effect of using the logarithm is that we can interpret Equation (24) in terms of the surprise value (also called information content) $-\log p_O(x_O)$ [36] of the pattern x_O and the surprise value of its parts with respect to any partition π. Rewriting Equation (22) 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 Equation (24) from this perspective we can then say that a pattern is an entity if the sum of the surprise values of its parts is larger than the surprise value of the whole.
- In coding theory, the Kraft-McMillan theorem [37] tells us that the optimal length (in a uniquely decodable binary code) of a codeword 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 codeword length and the chosen codeword length is

$$-\log q(x) - (-\log p(x)) = \log \frac{p(x)}{q(x)}. \tag{27}$$

  Then we can interpret the specific local integration as a difference in codeword 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 pattern) 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 codeword 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 codeword is shorter than all possible product code codewords. This means there is no partition with respect to which the product code for the pattern x_O has a shorter codeword than the joint code. So ι-entities are patterns that are shorter to encode with the joint code than a product code. Patterns that have a shorter codeword in a product code associated to a partition π have negative SLI with respect to this π and are therefore not ι-entities.
- 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 $\mathrm{MI}_\pi(X_O)$ [9,10] with respect to π, i.e.,

$$\mathrm{MI}_\pi(X_O) := \sum_{x_O\in\mathcal{X}_O} p_O(x_O) \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)} \tag{28}$$

$$= \sum_{x_O\in\mathcal{X}_O} p_O(x_O)\, \mathrm{mi}_\pi(x_O). \tag{29}$$

  The multi-information plays a role in measures of complexity and information integration [35]. The generalisation from bipartitions to arbitrary partitions is applied to expectation values similar to the multi-information above in Tononi [38]. The relations of our localised measure (in the sense of [11]) 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 patterns since they are properties of the

random variables X_O and not of patterns x_O. We also show in Corollary 2 that if x_O is an ι-entity that X_O (the joint random variable) has a positive $\mathrm{MI}_\pi(X_O)$ for all partitions π and is therefore a set of "integrated" random variables.

3.3. Properties of Specific Local Integration

This section investigates the specific local integration (SLI) (see Definition 5). After giving its expression for deterministic systems it proves upper bounds constructively and constructs an example of negative SLI.

3.3.1. Deterministic Case

Theorem 2 (Deterministic specific local integration)**.** *Given a deterministic Bayesian network (Definition A10), a uniform initial distribution over X_{V_0} (V_0 is the set of nodes without parents), and a pattern x_O with $O \subseteq V$ the SLI of x_O with respect to partition π can be expressed more specifically: Let $N(x_O)$ refer to the number of trajectories in which x_O occurs. Then*

$$\mathrm{mi}_\pi(x_O) = (|\pi| - 1)\log|\mathcal{X}_{V_0}| + \log N(x_O) - \sum_{b\in\pi}\log N(x_b). \tag{30}$$

Proof. See Appendix C.2. □

The first term in Equation (30) is always positive if the partition and the set of random variables are not trivial (i.e., have cardinality larger than one) and is a constant for partitions of a given cardinality. The second term is also always non-negative for patterns x_O that actually occur in the system and rises with the number of trajectories that lead to it. The third term is always non-positive and becomes more and more negative the higher the number of trajectories that lead to the parts of the pattern occurring.

This shows that to maximise SLI for fixed partition cardinality we need to find patterns that have a high number of trajectories leading to them and a low number of occurrences for all their parts. Since the number of occurrences of the parts cannot be lower than the number of occurrences of the whole, we should get a maximum SLI for patterns whose parts occur only if the whole occurs. This turns out to be true also for the non-deterministic systems as we prove in Theorem 4.

Conversely, if we can increase the number of occurrences of the parts of the pattern without increasing the occurrences of the whole pattern occurring we minimise the SLI. This leads to the intuition that as often as possible as many parts as possible (i.e., all but one) should co-occur without the whole pattern occurring. This consistently leads to negative SLI as we will show for the non-deterministic case in Theorem 5.

3.3.2. Upper Bounds

In this section we present the upper bounds of SLI. These are of general interest, but the constructive proof also provides an intuition for what kind of patterns have large SLI.

We first show constructively that if we can choose the Bayesian network and the pattern then SLI can be arbitrary large. This construction sets the probabilities of all blocks equal to the probability of the pattern and implies that each of the parts of the pattern occurs only if the entire pattern occurs. The simplest example is one binary random variable determining another to always be in the same state, then the two patterns with both variables equal have this property. 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. A simple extension of this example is used in the proof of the least upper bound. First we prove that there are Bayesian networks that achieve a particular SLI value. This will be used in the proofs that follow. For this we first define the anti-patterns which are patterns that differ to a given pattern at every random variable that is specified.

Definition 9 (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:*

$$\neg(x_O) := \{\bar{x}_O \in \mathcal{X}_O : \forall i \in O, \bar{x}_i \neq x_i\}. \tag{31}$$

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.

Theorem 3 (Construction of a pattern with maximum SLI)**.** *Given a probability $q \in (0,1)$ and a positive natural number n there is a Bayesian network $\{X_i\}_{i\in V}$ with $|V| \geq n$ and a pattern x_O such that*

$$\mathrm{mi}_\pi(x_O) = -(n-1)\log q. \tag{32}$$

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 $V \setminus O$ is arbitrary:

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

$$\begin{aligned}
\mathrm{mi}_\pi(x_O) &= \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)} && (33)\\
&= \log \frac{p_O(x_O)}{\prod_{b\in\pi} \sum_{\bar{x}_{O\setminus b}} p_O(x_b, \bar{x}_{O\setminus b})} && (34)\\
&= \log \frac{p_O(x_O)}{\prod_{b\in\pi} \left(p_O(x_O) + \sum_{\bar{x}_{O\setminus b} \neq x_{O\setminus b}} p_O(x_b, \bar{x}_{O\setminus b})\right)} && (35)\\
&= \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_O(x_O)} && (36)\\
&= \log \frac{p_O(x_O)}{p_O(x_O)^n} && (37)\\
&= -(n-1)\log q. && (38)
\end{aligned}$$

□

Theorem 4 (Upper bound of SLI). *For any Bayesian network $\{X\}_{i\in V}$ and pattern x_O with fixed $p_O(x_O) = q$*

1. *The tight upper bound of the SLI with respect to any partition π with $|\pi| = n$ fixed is*

$$\max_{\{\{X_i\}_{i\in V}:\exists x_O, p_O(x_O)=q\}} \max_{\{\pi:|\pi|=n\}} \mathrm{mi}_\pi(x_O) \leq -(n-1)\log q. \tag{39}$$

2. *The upper bound is achieved if and only if for all $b \in \pi$ we have*

$$p_b(x_b) = p_O(x_O) = q. \tag{40}$$

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

Proof. **ad 1** By Definition 5 we have

$$\mathrm{mi}_\pi(x_O) = \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)}. \tag{41}$$

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

$$\begin{aligned} p_b(x_b) &= \sum_{\bar{x}_{O\setminus b}} p_O(x_b, \bar{x}_{O\setminus b}) && (42)\\ &= p_O(x_O) + \sum_{\bar{x}_{O\setminus b}\neq x_{O\setminus b}} p_O(x_b, \bar{x}_{O\setminus b}) && (43)\\ &\geq p_O(x_O). && (44) \end{aligned}$$

Plugging this into Equation (41) for every $p_b(x_b)$ we get

$$\begin{aligned} \mathrm{mi}_\pi(x_O) &= \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)} && (45)\\ &\leq \log \frac{p_O(x_O)}{p_O(x_O)^{|\pi|}} && (46)\\ &= -(|\pi|-1)\log p_O(x_O). && (47) \end{aligned}$$

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 patterns x_O such that this upper bound is achieved. The construction of such a Bayesian network and a pattern x_O was presented in Theorem 3.

ad 2 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

$$\begin{aligned} & \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)} = -(|\pi|-1)\log p_O(x_O) && (48)\\ \Leftrightarrow \quad & \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)} = \log \frac{p_O(x_O)}{p_O(x_O)^{|\pi|}} && (49)\\ \Leftrightarrow \quad & \prod_{b\in\pi} p_b(x_b) = p_O(x_O)^{|\pi|}, && (50) \end{aligned}$$

and because $p_b(x_b) \geq p_O(x_O)$ (Equation (44)) 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 3 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 a pattern $(x_b, \bar{x}_{O\setminus b})$ with $\bar{x}_{O\setminus b} \neq x_{O\setminus b}$ i.e., $p_O(x_b, \bar{x}_{O\setminus b}) > 0$. Recalling Equation (43) we then see that

$$p_b(x_b) = p_O(x_O) + \sum_{\bar{x}_{O\setminus b} \neq x_{O\setminus b}} p_O(x_b, \bar{x}_{O\setminus b}) \tag{51}$$

$$> p_O(x_O). \tag{52}$$

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 pattern that achieves this bound.
- The least upper bound of SLI increases with the improbability of the pattern and the number of parts that it is split into. If $p_O(x_O) \to 0$ then we can have $\mathrm{mi}_\pi(x_O) \to \infty$.
- Using this least upper bound it is easy to see the least upper bound for the SLI of a pattern x_O across all partitions $|\pi|$. We just have to note that $|\pi| \leq |O|$.
- Since it is the minimum value of SLI with respect to arbitrary partitions the least upper bound of SLI is also an upper bound for CLI. It may not be the least upper bound however.

3.3.3. Negative SLI

This section shows that SLI of a pattern x_O with respect to partition π can be negative *independently* of 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). The construction which achieves this also serves as an example of patterns with low SLI. We conjecture that this construction might provide the greatest lower bound but have not been able to prove this yet. An intuitive description of the construction is that patterns which either occur as a whole or missing exactly one part always have negative SLI.

Theorem 5. *For any given probability $q < 1$ and cardinality $|\pi| = n > 1$ of a partition π there exists a Bayesian network $\{X_i\}_{i\in V}$ with a pattern x_O such that $q = p_O(x_O)$ and*

$$\mathrm{mi}_\pi(x_O) = \log \frac{q}{\left(1 - \frac{1-q}{n}\right)^n} < 0. \tag{53}$$

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 and Definition 9 for $\neg(x_A)$):

1. for all $i \in O$ let $|\mathcal{X}_i| = n$
2. for every block $b \in \pi$ let $|b| = \frac{|O|}{|\pi|}$,
3. for $\bar{x}_O \in \mathcal{X}_O$ let:

$$p_O(\bar{x}_O) := \begin{cases} q & \text{if } \bar{x}_O = x_O, \\ \frac{1-q-d}{\sum_{b\in\pi} |\neg(x_b)|} & \text{if } \exists c \in \pi \text{ s.t. } \bar{x}_{O\setminus c} = x_{O\setminus c} \wedge \bar{x}_c \neq x_c, \\ \frac{d}{|\neg(x_O)|} & \text{if } \bar{x}_O \in \neg(x_O), \\ 0 & \text{else.} \end{cases} \tag{54}$$

Here d parameterises the probability of any pattern in $\neg(x_O)$ occurring. We will carry it through the calculation but then end up setting it to zero.

Next we calculate the SLI. First note that, according to 1. and 2., 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 3, for all $b \in \pi$

$$\sum_{\bar{x}_{O\setminus b} \neq x_{O\setminus b}} p_O(x_b, \bar{x}_{O\setminus b}) = \sum_{c\in\pi\setminus b} \sum_{\bar{x}_c \neq x_c} p_O(x_b, x_{O\setminus(b\cup c)}, \bar{x}_c) \tag{55}$$

$$= \sum_{c\in\pi\setminus b} \sum_{\bar{x}_c \neq x_c} \frac{1-q-d}{\sum_{b\in\pi} |\neg(x_b)|} \tag{56}$$

$$= \sum_{c\in\pi\setminus b} \sum_{\bar{x}_c \neq x_c} \frac{1-q-d}{m|\pi|} \tag{57}$$

$$= \sum_{c\in\pi\setminus b} \frac{1-q-d}{m|\pi|} |\neg(x_c)| \tag{58}$$

$$= \frac{|\pi|-1}{|\pi|}(1-q-d) \tag{59}$$

Plug this into the SLI definition:

$$\mathrm{mi}_\pi(x_O) = \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)} \tag{60}$$

$$= \log \frac{q}{\prod_{b\in\pi} q + \sum_{\bar{x}_{O\setminus b} \neq x_{O\setminus b}} p_O(x_b, \bar{x}_{O\setminus b})} \tag{61}$$

$$= \log \frac{q}{\prod_{b\in\pi} q + \frac{|\pi|-1}{|\pi|}(1-q-d)} \tag{62}$$

$$= \log \frac{q}{\left(q + \frac{|\pi|-1}{|\pi|}(1-q-d)\right)^{|\pi|}} \tag{63}$$

If we now set $d = 0$ we get:

$$\mathrm{mi}_\pi(x_O) = \log \frac{q}{\left(1 - \frac{1-q}{|\pi|}\right)^{|\pi|}}. \tag{64}$$

Then we can use Bernoulli's inequality (The authors thank von Eitzen [39] for pointing this out. An example reference for Bernoulli's inequality is Bullen [40]). to prove that this is negative for $0 < q < 1$ and $|\pi| \geq 2$. Bernoulli's inequality is

$$(1+x)^n \geq 1 + nx \tag{65}$$

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

$$\left(1 - \frac{1-q}{|\pi|}\right)^{|\pi|} > q \tag{66}$$

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

Remarks:

- The achieved value in Equation (53) is also our best candidate for a greatest lower bound of SLI for 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 the probability q of the whole pattern occurring is chosen) to the patterns $\bar{x}_O$ that are *almost* the same as the pattern

x_O. These are almost the same in a precise sense: They differ in exactly one of the blocks of π, i.e., they differ by as little as can possibly be resolved/revealed by the partition π.
- In order to achieve the negative SLI of Equation (64) the requirement is only that Equation (59) 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 3. of the proof such that Equation (59) still holds. This is not necessary for our purpose here as we only want to show the existence of patterns achieving the negative value.
- Since it is the minimum value of SLI with respect to arbitrary partitions the candidate for the greatest lower bound of SLI is also a candidate for the greatest lower bound of CLI.

3.4. 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 paper. 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 pattern is a block in such a finest partition leading to a particular value of SLI. The theorem therefore reveals that ι-entities can not only be motivated heuristically as we tried to do in Section 3.2 but in fact play a special role within the trajectories they occur in. Furthermore, this theorem allows additional interpretations of the ι-entities which will be discussed in Section 3.5.

The main tool we use for the proof, the disintegration hierarchy and especially its refinement free version are also interesting structure in their own right since they define a hierarchy among the partitions of trajectories that we did not anticipate. In the case of the refinement free version the disintegration theorem tells us that this hierarchy among partitions of trajectories turns out to be a hierarchy of splits of the trajectory into *ci*-entities.

Definition 10 (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:*

1.
$$\mathfrak{D}_1(x_V) := \underset{\pi\in\mathfrak{L}(V)}{\arg\min}\, \mathrm{mi}_\pi(x_V) \tag{67}$$

2. and for $i > 1$:
$$\mathfrak{D}_i(x_V) := \underset{\pi\in\mathfrak{L}(V)\setminus\mathfrak{D}_{\prec i}(x_V)}{\arg\min}\, \mathrm{mi}_\pi(x_V). \tag{68}$$

where $\mathfrak{D}_{\prec i}(x_V) := \bigcup_{j<i}\mathfrak{D}_j(x_V)$. 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 if there is more than one.
- 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 11. *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*

$$\mathfrak{E}_{\lhd\pi} := \{\xi \in \mathfrak{E} : \xi \lhd \pi\}. \tag{69}$$

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

Definition 12 (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:*

1.
$$\mathfrak{D}_1^{\blacktriangleleft}(x_V) := \{\pi \in \mathfrak{D}_1(x_V) : \mathfrak{D}_1(x_V)_{\lhd\pi} = \emptyset\}, \tag{70}$$

2. *and for $i > 1$:*
$$\mathfrak{D}_i^{\blacktriangleleft}(x_V) := \{\pi \in \mathfrak{D}_i(x_V) : \mathfrak{D}_{\prec i}(x_V)_{\lhd\pi} = \emptyset\} \tag{71}$$

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 6 (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.*

1. *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:*
   (a) *b is a singleton, i.e., $b = \{i\}$ for some $i \in V$, or*
   (b) *x_b is completely locally integrated, i.e., $\iota(x_b) > 0$.*
2. *Conversely, for any completely locally integrated pattern 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 1 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 $\mathrm{mi}_\xi(x_b) \leq 0$. 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 $p_b(x_b) \leq \prod_{d\in\xi} p_d(x_d)$ with $|\xi| > 1$. However, then

$$\mathrm{mi}_\pi(x_V) = \log \frac{p_V(x_V)}{p_b(x_b)\prod_{c\in\pi\setminus b} p_c(x_c)} \tag{72}$$

$$\geq \log \frac{p_V(x_V)}{\prod_{d\in\xi} p_d(x_d)\prod_{c\in\pi\setminus b} p_c(x_c)} \tag{73}$$

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

$$\mathrm{mi}_\pi(x_V) = \log \frac{p_V(x_V)}{\prod_{d\in\xi} p_d(x_d)\prod_{c\in\pi\setminus b} p_c(x_c)}. \tag{74}$$

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

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

Second, let

$$\mathrm{mi}_\pi(x_V) > \log \frac{p_V(x_V)}{\prod_{d\in\xi} p_d(x_d) \prod_{c\in\pi\setminus b} p_c(x_c)}. \tag{75}$$

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

$$\mathrm{mi}_\rho(x_V) < \mathrm{mi}_\pi(x_V), \tag{76}$$

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

ad 2 By assumption x_A is completely locally integrated. Then 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 $k \geq i^A$. 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

$$\mathrm{mi}_{\xi|_A}(x_A) = \log \frac{p_A(x_A)}{\prod_{c\in\xi|_A} p_c(x_c)} > 0 \tag{77}$$

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

$$\mathrm{mi}_\xi(x_V) = \log \frac{p_V(x_V)}{\prod_{c\in\xi|_A} p_c(x_c) \prod_{i\in V\setminus A} p_i(x_i)} \tag{78}$$

$$> \log \frac{p_V(x_V)}{p_A(x_A) \prod_{i\in V\setminus A} p_i(x_i)} \tag{79}$$

$$= \mathrm{mi}_{\pi^A}(x_V). \tag{80}$$

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 $\mathfrak{D}_{\prec i^A}(x_V)_{\lhd \pi^A} = \emptyset$ and $\pi^A \in \mathfrak{D}_{i^A}^{\blacktriangleleft}(x_V)$.

□

We mentioned in Section 3.2 that the expectation value of SLI $\mathrm{mi}_\pi(x_A)$ is the (specific) multi-information $\mathrm{MI}_\pi(X_A)$. A positive SLI value of x_A implies a positive expectation value $\mathrm{MI}_\pi(X_A)$. Therefore every ι-entity x_A implies positive specific multi-informations $\mathrm{MI}_\pi(X_A)$ with respect to any partition π. We put this into the following corollary.

Corollary 2. *Under the conditions of Theorem 6 and 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 (not only locally) integrated, i.e., $\mathrm{I}(X_b) > 0$.*

here

$$\mathrm{I}(X_A) := \min_{\pi\in\mathfrak{L}(A)\setminus \mathbf{0}} \mathrm{MI}_\pi(X_A). \tag{81}$$

Proof. Since $\mathrm{MI}_\pi(X_A)$ is a Kullback–Leibler divergence we know from Gibbs' inequality that $\mathrm{MI}_\pi(X_A) \geq 0$ and $\mathrm{MI}_\pi(X_A) = 0$ if and only if for all $x_A \in \mathcal{X}_A$ we have $p_A(x_A) = \prod_{b\in\pi} p_b(x_b)$. To see that $\mathrm{MI}_\pi(X_A)$ is a Kullback–Leibler divergence note:

$$\mathrm{MI}_\pi(X_A) := \sum_{x_A\in\mathcal{X}_A} p_A(x_A)\,\mathrm{mi}_\pi(x_A) \tag{82}$$

$$= \sum_{x_A\in\mathcal{X}_A} p_A(x_A)\log\frac{p_A(x_A)}{\prod_b p_b(x_b)} \tag{83}$$

$$= \mathrm{KL}[p_A||\prod_{b\in\pi} p_b]. \tag{84}$$

Now let a specific $x_A \in \mathcal{X}_A$ be a ι-entity. Then for all $\pi \in \mathfrak{L}(A)\setminus \mathbf{0}$ we have

$$\log\frac{p_A(x_A)}{\prod_b p_b(x_b)} > 0, \tag{85}$$

which implies that

$$p_A(x_A) \neq \prod_b p_b(x_b) \tag{86}$$

and therefore

$$\mathrm{KL}[p_A||\prod_{b\in\pi} p_b] > 0 \tag{87}$$

which implies $\mathrm{I}(X_A) > 0$. □

3.5. Disintegration Interpretation

In Section 3.2 we motivated our choice of positive complete local integration as a criterion for entities. This motivation is purely heuristic and starts from the intuition that an entity is a structure for which every part makes every other part more probable. While this heuristic argument seems sufficiently intuitive to be of a certain value we would much rather have a formal reason why an entity criterion is a "good" entity criterion. In other words we would ideally have a formal problem that is best solved by the entities satisfying the criterion. An example of a measure that has such an associated interpretation is the mutual information whose maximum over the input distributions is the channel capacity. Without a formal problem associated to ι-entities there remains a risk that they (and maybe the whole concept of entities and identity over time) are artefacts of an ill-conceived conceptual approach.

Currently, we are not aware of an analogous formal problem that is solved by ι-entities. However, the different viewpoint provided by the disintegration theorem may be a first step towards finding such a problem. We will now discuss some alternative interpretations of SLI and see how CLI can be seen from a different perspective due to the disintegration theorem. These interpretations also exhibit why we chose to include the logarithm into the definition of SLI.

Using the disintegration theorem (Theorem 6) allows us to take another point of view on ι-entities. 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 partitions whose blocks are completely integrated patterns i.e., they only contain ι-entities. At the same time the blocks of all those partitions together are *all* ι-entities that occur in that trajectory.

A partition in the refinement-free disintegration hierarchy is always a minimal/finest partition 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 codeword 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.,

$$\mathrm{mi}_\pi(x_V) = \log \frac{p_V(x_V)}{\prod_{c\in\pi} p_c(x_c)} \tag{88}$$

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 codeword 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 codewords. 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 codeword length* of the product code:

$$\sum_{x_V\in\mathcal{X}_V} p_V(x_V)(-\log \prod_{c\in\pi} p_c(x_c)) \tag{89}$$

which is the more important measure for encoding in general, might not be short at all, i.e., it might not be an efficient code for arbitrary trajectories. The product codes based on partitions in $\mathfrak{D}_1^{\blacktriangleleft}(x_V)$ are specifically adapted to assign a short codeword to x_V, i.e., to a single trajectory or story of this system. As product codes 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.

Nonetheless, the product codes exist (they can be generated using Huffman coding or arithmetic coding [37] based on the product probability) and are uniquely decodable. The parts/blocks of them are the ι-entities. We mentioned before that we would like to find a problem that is solved by ι-entities. This is then equivalent to finding a problem that is solved by the according product codes. Can we construct such a problem? This question is still open. A possible direction for finding such a problem may be the following line of reasoning. 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., say 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 codeword 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.

Speculating about where the two conditions may arise in an actual problem, 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/happened in this trajectory needs to be communicated more often than what happens/happened 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., [41]). So the entities that mankind has partitioned its history into might be related to the ι-entities of the universe's history. These would compose the shortest product codes for what actually happened. The disintegration level might be chosen to optimise rates of model learning.

Recall that this kind of product code is not the optimal code in general (which would be the one with shortest expected codeword 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 codeword lengths reflecting a desire to communicate efficiently about all counterfactual possibilities as well.

3.6. Related Approaches

We now discuss in some more detail than in Section 1.3 the approaches of Beer [14] and Balduzzi [26].

In Beer [14] 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 (or temporal part) of a pattern to the next. The used example patterns are the famous block, blinker, and glider and they 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*. Organisational closure occurs if the same process-component relations reoccur at a later time. Boundaries of the spatiotemporal patterns are identified by determining the cells around the pattern that have to be fixed to get re-occurrence of the organisation.

Beer [14] 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 using the entity criterion we propose does not need similar choices to be made since it is not based on the re-occurrence of any organisation. Later time-slices of ι-entities need no organisational (or any other) similarity to earlier ones. Another, possibly only small, advantage is that our criterion is formalised and reasonably simply to state. Whether this is possible for the organisational closure based entities remains to be seen.

This is related to the philosophical discussion about identity across possible worlds [33].

Some further parallels can be drawn between the present work and Balduzzi [26] especially if we take into account the disintegration theorem. Given a trajectory (entire time-evolution) of the system in both cases a partition is sought which fulfills a particular trajectory-wide optimality criterion. Also in both cases, each block of the trajectory-wide partition fulfills a condition with respect to its own partitions. For our conditions the disintegration theorem exposes the direct connection between the trajectory-wide and the block-specific conditions. Such a connection is not known for other approaches. The main reason for this might be the simpler formal expression of CLI and SLI compared to the IIT approaches.

In how far our approach and the IIT approaches lead to coinciding or contradicting results is beyond the scope of this paper and constitutes future work. One avenue to pursue here are differences with respect to entities occurring in multiple trajectories as well as the possibility of overlapping entities within single trajectories.

4. Examples

In this section we investigate the structure of integrated and completely locally integrated spatiotemporal patterns as it is revealed by the disintegration hierarchy. First we take a quick look at the trivial case of a set of independent random variables. Then we look at two very simple multivariate Markov chains. We use the disintegration theorem (Theorem 6) to extract the completely locally integrated spatiotemporal patterns.

4.1. Set of Independent Random Variables

Let us first look at a set $\{X_i\}_{i\in V}$ of independently and identically distributed random variables. For each trajectory $x_V \in \mathcal{X}_V$ we can then calculate SLI with respect to a partition $\pi \in \mathfrak{L}(V)$. For every $A \subseteq V$ and every $x_A \in \mathcal{X}_A$ we have $p_A(x_A) = \prod_{i\in A} p_i(x_i)$. Then we find for every $\pi \in \mathfrak{L}(V)$:

$$\mathrm{mi}_\pi(x_V) = 0. \tag{90}$$

This shows that the disintegration hierarchy for each $x_V \in \mathcal{X}_V$ contains only a single disintegration level $\mathfrak{D}(x_V) = \{\mathfrak{D}_1\}$ with $\mathfrak{D}_1 = \mathfrak{L}(V)$. The finest partition of $\mathfrak{L}(V)$ is its zero element $\mathbf{0}$ which then constitutes the only element of the refinement-free disintegration level $\mathfrak{D}_1^{\blacktriangleleft} = \{\mathbf{0}\}$. Recall that the zero element of a partition lattice only consists of singleton sets as blocks. The set of completely locally integrated patterns i.e., the set of ι-entities in a given trajectory x_V is then the set $\{x_i : i \in V\}$.

Next we will look at more structured systems.

4.2. Two Constant and Independent Binary Random Variables: $MC^=$

4.2.1. 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

$$\mathrm{pa}(j,t) = \begin{cases} \emptyset & \text{if } t = 0, \\ \{(j,t-1)\} & \text{else,} \end{cases} \tag{91}$$

$$p_{j,t}(x_{j,t}|x_{j,t-1}) = \delta_{x_{j,t-1}}(x_{j,t}) = \begin{cases} 1 & \text{if } x_{j,t} = x_{j,t-1}, \\ 0 & \text{else,} \end{cases} \tag{92}$$

$$p_{j,0}(x_{j,0}) = 1/4. \tag{93}$$

The Bayesian network can be seen in Figure 4.

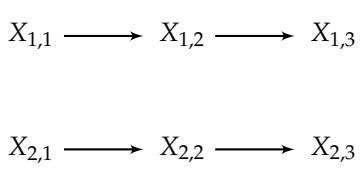


Figure 4. Bayesian network of $MC^=$. There is no interaction between the two processes.

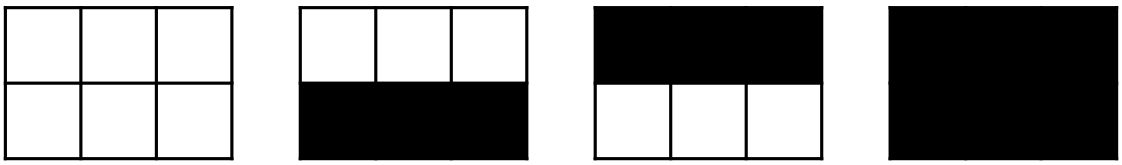
4.2.2. Trajectories

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:

$$x_V = (x_{1,0}, x_{2,0}, x_{1,1}, x_{2,1}, x_{1,2}, x_{2,2}) = \begin{cases} (0,0,0,0,0,0) & \text{if } x_{1,0} = 0, x_{2,0} = 0; \\ (0,1,0,1,0,1) & \text{if } x_{1,0} = 0, x_{2,0} = 1; \\ (1,0,1,0,1,0) & \text{if } x_{1,0} = 1, x_{2,0} = 0; \\ (1,1,1,1,1,1) & \text{if } x_{1,0} = 1, x_{2,0} = 1. \end{cases} \tag{94}$$

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 Figure 5.



(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 5. Visualisation of the four possible trajectories of $MC^{=}$. 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.

4.2.3. Partitions of Trajectories

The disintegration hierarchy is composed out of all partitions in the lattice of partitions $\mathfrak{L}(V)$. Recall 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, spatiotemporally and not only spatially extended patterns.

The number of partitions $|\mathfrak{L}(V)|$ of a set of $|V| = 6$ elements is $\mathcal{B}_6 = 203$ ($\mathcal{B}_n$ is the Bell number of n). 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 Stirling number $\mathcal{S}(|V|, |\pi|)$. For $|V| = 6$ we find the Stirling numbers:

$\lvert\pi\rvert$	1	2	3	4	5	6
$\mathcal{S}(\lvert V\rvert, \lvert\pi\rvert)$	1	31	90	65	15	1

(95)

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.

4.2.4. 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 5:

$$\mathrm{mi}_\pi(x_V) := \log \frac{p_V(x_V)}{\prod_{b\in\pi} p_b(x_b)}. \tag{96}$$

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 Figure 6.

We can see in Figure 6 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.

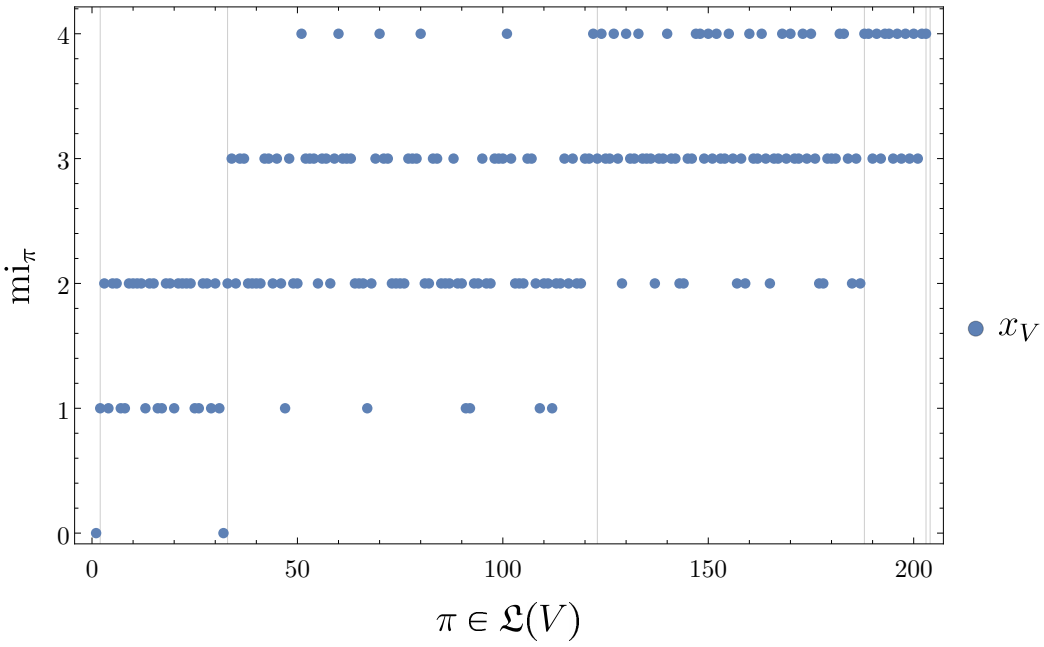


Figure 6. Specific local integrations $\mathrm{mi}_\pi(x_V)$ of any of the four trajectories x_V seen in Figure 5 with respect to all $\pi \in \mathfrak{L}(V)$. The partitions are ordered according to an enumeration with increasing cardinality $|\pi|$ ((see Pemmaraju and Skiena [42], Chapter 4.3.3) for the method). We indicate with vertical lines at what partitions the cardinality $|\pi|$ increases by one.

4.2.5. Disintegration Hierarchy

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 10) and look at the Hasse diagrams of each of the disintegration levels $\mathfrak{D}_i(x_V)$ partially ordered by refinement. If we sort the partitions of any trajectory of $MC^=$ according to increasing SLI value we obtain Figure 7. There we see groups of partitions attaining the SLI values $\{0,1,2,3,4\}$ (precisely) 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

(97)

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 (see Appendix B for the definition) visualise the set of partitions in each disintegration level partially ordered by refinement $\lhd$. The Hasse diagrams are shown in Figure 8. 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 $\mathfrak{D}_2$ (Figure 8b) 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. As it requires further notational overhead and is straightforward we do not describe these symmetry properties formally. In order to see the symmetries, however, we visualise the partitions themselves in the Hasse diagrams in Figure 9. We also visualise examples of the different connected components in each disintegration level in Figure 10.

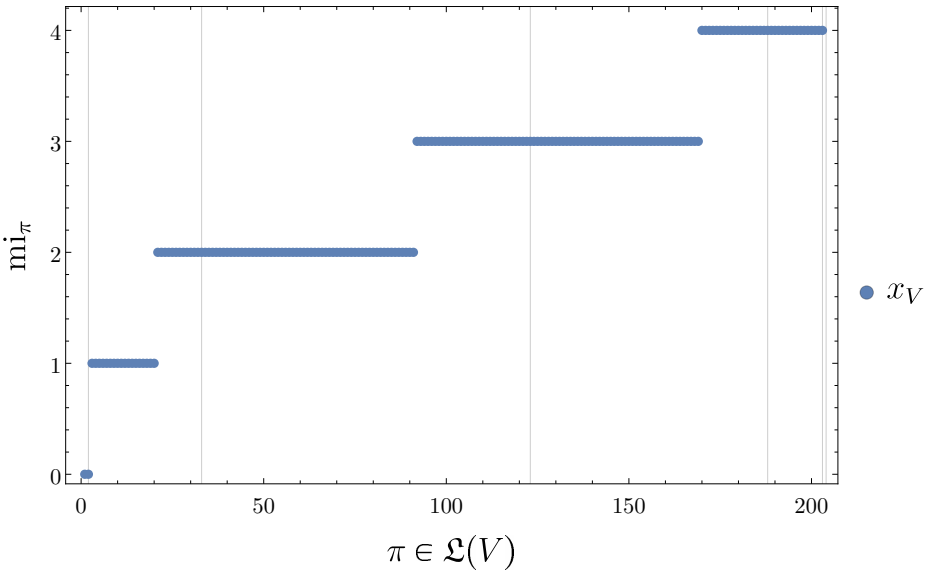
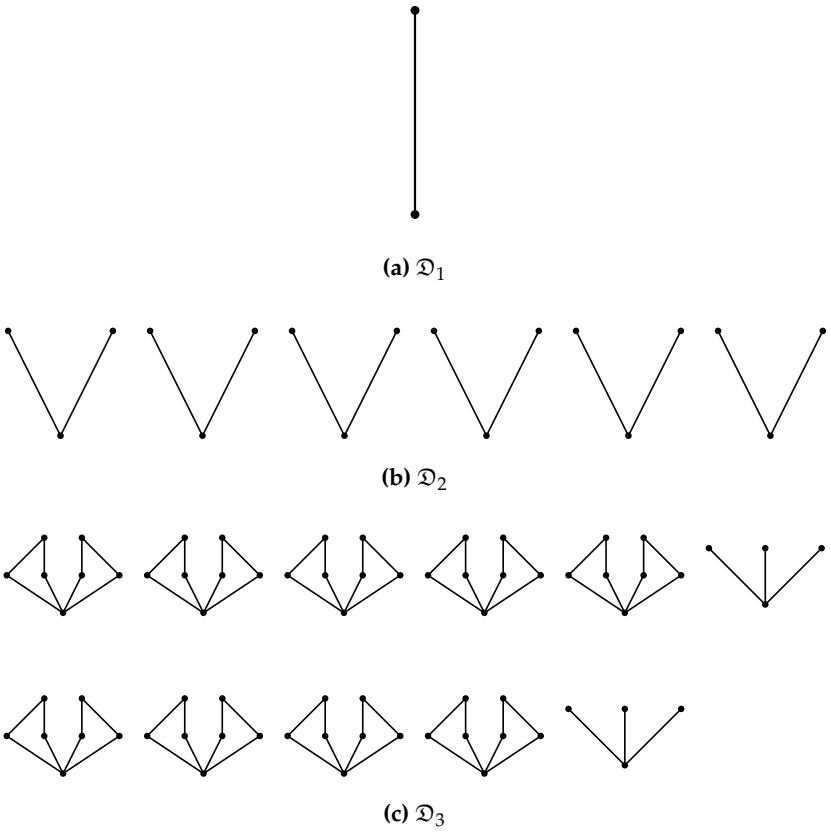


Figure 7. Same as Figure 6 but with the partitions sorted according to increasing SLI.

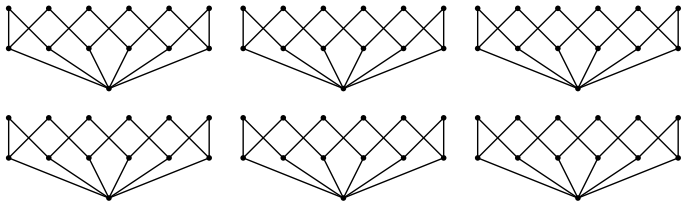


(a) $\mathfrak{D}_1$

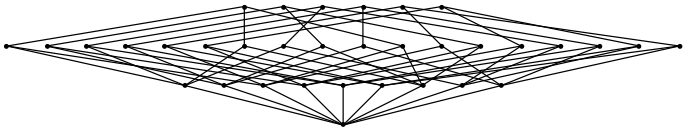
(b) $\mathfrak{D}_2$

(c) $\mathfrak{D}_3$

Figure 8. *Cont.*



(d) $\mathfrak{D}_4$



(e) $\mathfrak{D}_5$

Figure 8. 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.

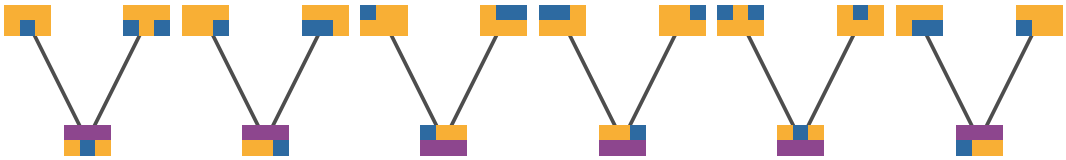


Figure 9. Hasse diagram of $\mathfrak{D}_2$ of $MC^{=}$ trajectories. Here we visualise the partitions at each vertex. The blocks of a partition are the cells of equal colour. Note that we can obtain all six disconnected components from one by symmetry operations that are respected by the joint probability distribution p_V. For example, we can shift each row individually to the left or right since every value is constant in each row. We can also switch top and bottom row since they have the same probability distributions even if 1 and 0 are exchanged.

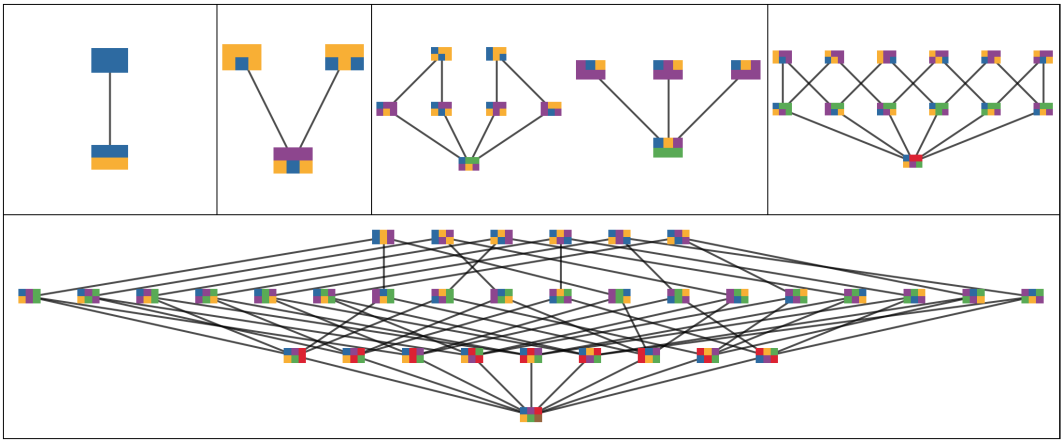


Figure 10. For each disintegration level of the trajectories of $MC^{=}$ we here show example connected components of Hasse diagrams with the partitions at each vertex visualised. The disintegration level increases clockwise from the top left. The blocks of a partition are the cells of equal colour.

Recall that due to the disintegration theorem (Theorem 6) 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 Figure 11.

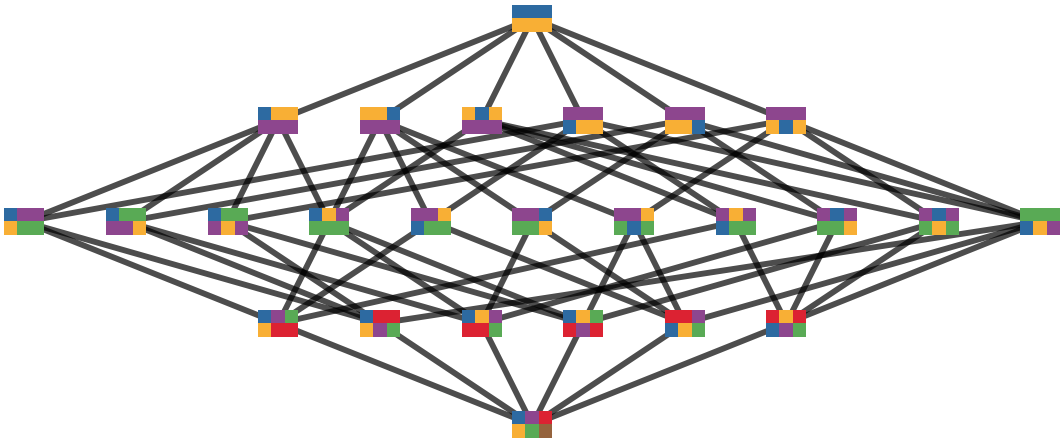


Figure 11. Hasse diagrams of the refinement-free disintegration hierarchy $\mathfrak{D}^{\blacktriangleleft}$ of $MC^{=}$ trajectories. Here we visualise the partitions at each vertex. The blocks of a partition are the cells of equal colour. It turns out that partitions that are on the same *horizontal* 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 Figure 9. To connect this to Figure 8 note that for each disintegration level $\mathcal{D}_i$ shown there as a Hasse diagram, the partitions on the i-th horizontal level (counting from the top) in the present figure are the minimal elements of that disintegration level.

4.2.6. Completely Integrated Patterns

Having looked at the disintegration hierarchy we now make use of it by extracting the completely (When it is clear from context that we are talking about complete local integration we drop "local" for the sake of readability.) integrated patterns (ι-entities) of the four trajectories of $MC^{=}$. Recall that due to the disintegration theorem (Theorem 6) we know that all blocks in partitions that occur in the refinement-free disintegration hierarchy are either singletons or correspond to ι-entities. If we look at the refinement-free disintegration hierarchy in Figure 11 we 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, which are trivially integrated as they cannot be partitioned, we end up with eight different blocks. Since the disintegration hierarchy is the same for all four possible trajectories these blocks are also the same for each of them (note that this is the case for $MC^{=}$ but not in general as we will see in Section 4.3). However, the patterns that result are different due to the different values within the blocks. We show the eight ι-entities and their complete local integration (Definition 6) on the first trajectory in Figure 12 and on the second trajectory in Figure 13.

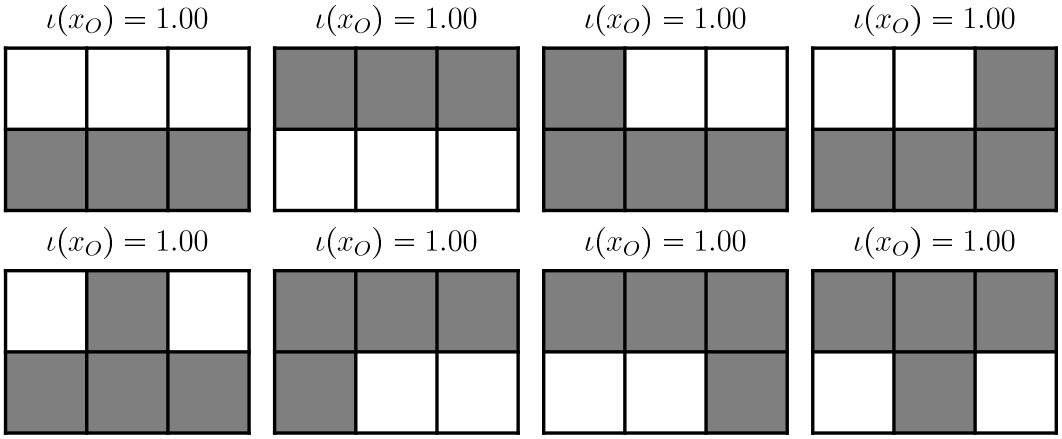


Figure 12. All distinct completely integrated composite patterns (singletons are not shown) on the first possible trajectory of $MC^{=}$. The value of complete local integration is indicated above each pattern. We display patterns by colouring the cells corresponding to random variables that are not fixed to any value by the pattern in grey. Cells corresponding to random variables that are fixed by the pattern are coloured according to the value i.e., white for 0 and black for 1.

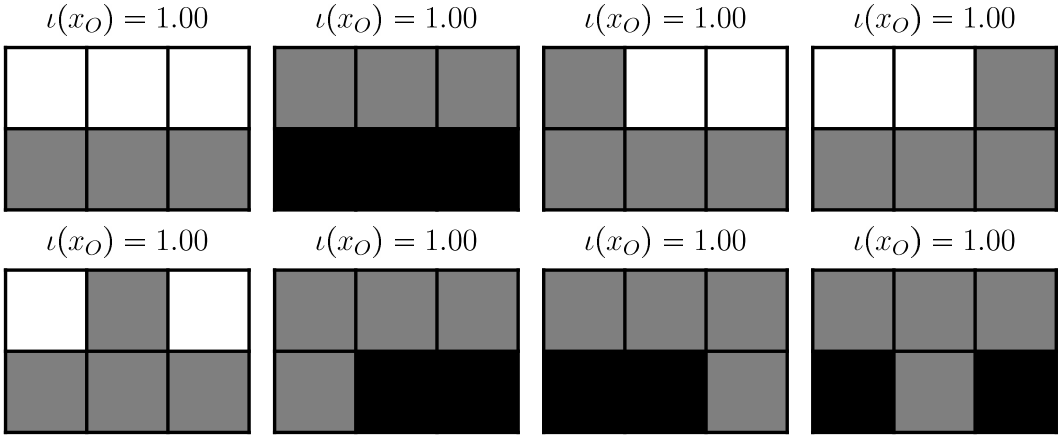


Figure 13. All distinct completely integrated composite patterns on the second possible trajectory of $MC^{=}$. The value of complete local integration is indicated above each pattern.

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 ι-entities. This is apparent when comparing Figures 12 and 13 and noting that each pattern occurring on the first trajectory has a corresponding pattern 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 pattern in Figure 12 there is a corresponding pattern in Figure 13 leaving the same cells grey.

If we are not interested in a particular trajectory, we can also look at all different ι-entities on any trajectory. For $MC^{=}$ these are shown in Figure 14.

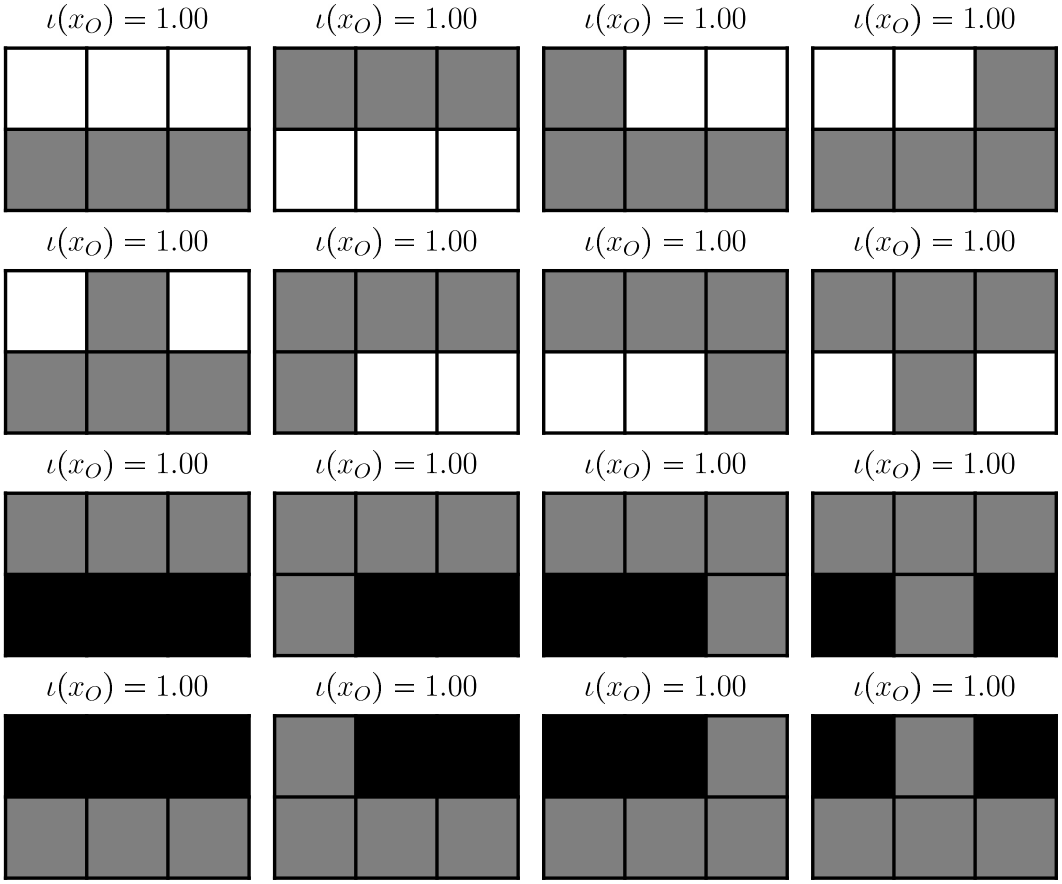


Figure 14. All distinct completely integrated composite patterns on all four possible trajectories of $MC^{=}$. The value of complete local integration is indicated above each pattern.

We see that all ι-entities 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 Equation (30) and noting that for $MC^{=}$ if any of the values $x_{j,t}$ is fixed by a pattern 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 ι-entities and all ι-entities x_O themselves are of the form $x_{j,S}$ with $S \subseteq T$. Let $N(x_{j,S}) =: N$ and plug this into Equation (30) for an arbitrary partition π:

$$\mathrm{mi}_\pi(x_O) = (|\pi| - 1)\log|\mathcal{X}_{V_0}| - \log\frac{\prod_{b\in\pi} N(x_b)}{N(x_O)} \tag{98}$$

$$= (|\pi| - 1)\log|\mathcal{X}_{V_0}| - \log\frac{N^{|\pi|}}{N} \tag{99}$$

$$= (|\pi| - 1)\log\frac{|\mathcal{X}_{V_0}|}{N}. \tag{100}$$

To get the complete local integration value we have to minimise this with respect to π where $|\pi| \geq 2$. So for $|\mathcal{X}_{V_0}| = 4$ and $N = 2$ we get $\iota(x_O) = 1$.

Another observation is that the ι-entities are all limited to one of the two rows. This shows on a simple example that, as we would expect, ι-entities cannot extend from one independent process to another.

4.3. Two Random Variables with Small Interactions

In this section we look at a system almost identical to that of Section 4.2 but with a kind of noise introduced. This allows all trajectories to occur and is designed to test whether the spatiotemporal patterns maintain integration in the face of noise.

4.3.1. Definition

We define the time- and space-homogeneous multivariate Markov chain MC^ϵ via the Markov matrix P with entries

$$P_{f(x_{1,t+1},x_{2,t+1}),f(x_{1,t},x_{2,t})} = p_{J,t+1}(x_{1,t+1}, x_{2,t+1}|x_{1,t}, x_{2,t}) \tag{101}$$

where we define the function $f : \{0,1\}^2 \to [1:4]$ via

$$f(0,0) = 1, f(0,1) = 2, f(1,0) = 3, f(1,1) = 4. \tag{102}$$

With this convention P is

$$P = \begin{pmatrix} 1-3\epsilon & \epsilon & \epsilon & \epsilon \\ \epsilon & 1-3\epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & 1-3\epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & 1-3\epsilon \end{pmatrix} \tag{103}$$

This means that the state of *both* random variables remains the same with probability $1-3\epsilon$ and transitions into each of the other possible combinations with probability ϵ. The noise then does not act independently on both random variables but disturbs the joint state. This makes ι-entities possible that extend across the two processes. In the following we set $\epsilon = 1/100$. The initial distribution is again the uniform distribution

$$p_{j,0}(x_{j,0}) = 1/4. \tag{104}$$

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 Figure 15.

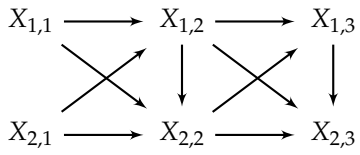
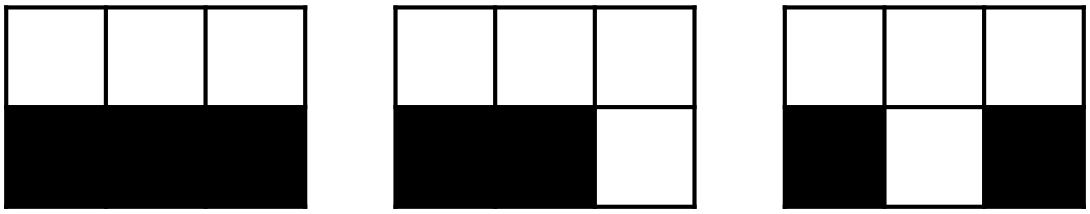


Figure 15. Bayesian network of MC^ϵ.

4.3.2. 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 Figure 16 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 16. 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 $t = 3$, and the third (in (**c**)) makes two.

4.3.3. 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 Figure 17.

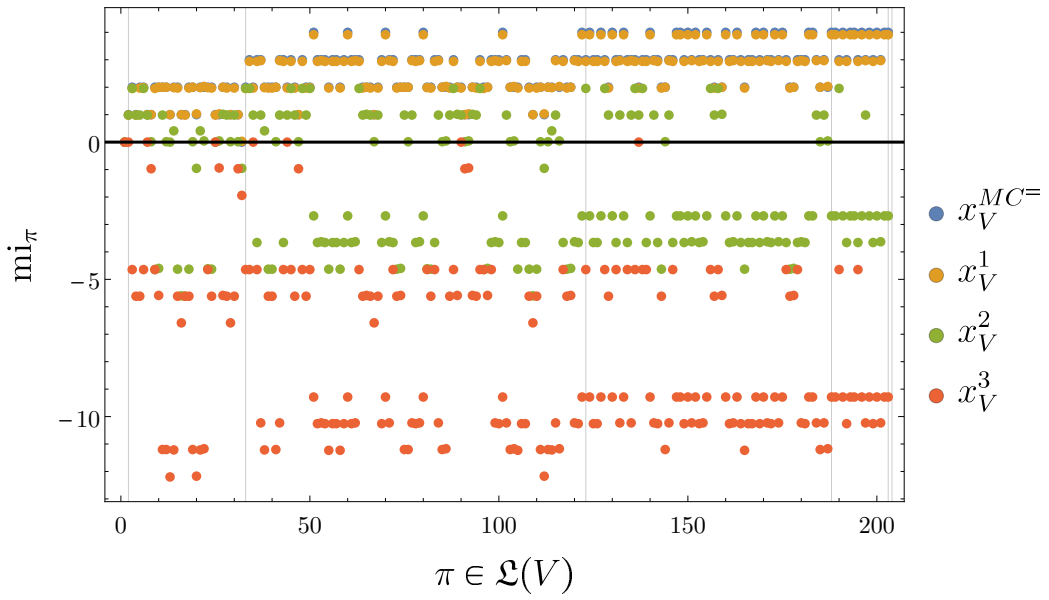


Figure 17. Specific local integrations $\mathrm{mi}_\pi(x_V)$ of one of the four trajectories of $MC^=$ (measured w.r.t. the probability distribution of $MC^=$), here denoted $x_V^{MC^=}$ (this is the same data as in Figure 6), and the three representative trajectories $x_V^k, x \in \{1,2,3\}$ of MC^ϵ (measured w.r.t. the probability distribution of MC^ϵ) seen in Figure 16 with respect to all $\pi \in \mathfrak{L}(V)$. The partitions are ordered as in Figure 6 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.

It turns out that the SLI values of x_V^1 are almost the same as those of $MC^=$ in Figure 6 with small deviations due to the noise. This should be expected as x_V^1 is also a possible trajectory of $MC^=$. 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.

4.3.4. Completely Integrated Patterns

In this section we look at the ι-entities 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 Figures 18–20. In contrast to the situation of $MC^{=}$ we now have ι-entities with varying values of complete local integration.

On the first trajectory x_V^1 we find all the eight patterns that are completely locally integrated in $MC^{=}$ (see Figure 13). These are also more than an order of magnitude more integrated than the rest of the ι-entities. This is also true for the other two trajectories.

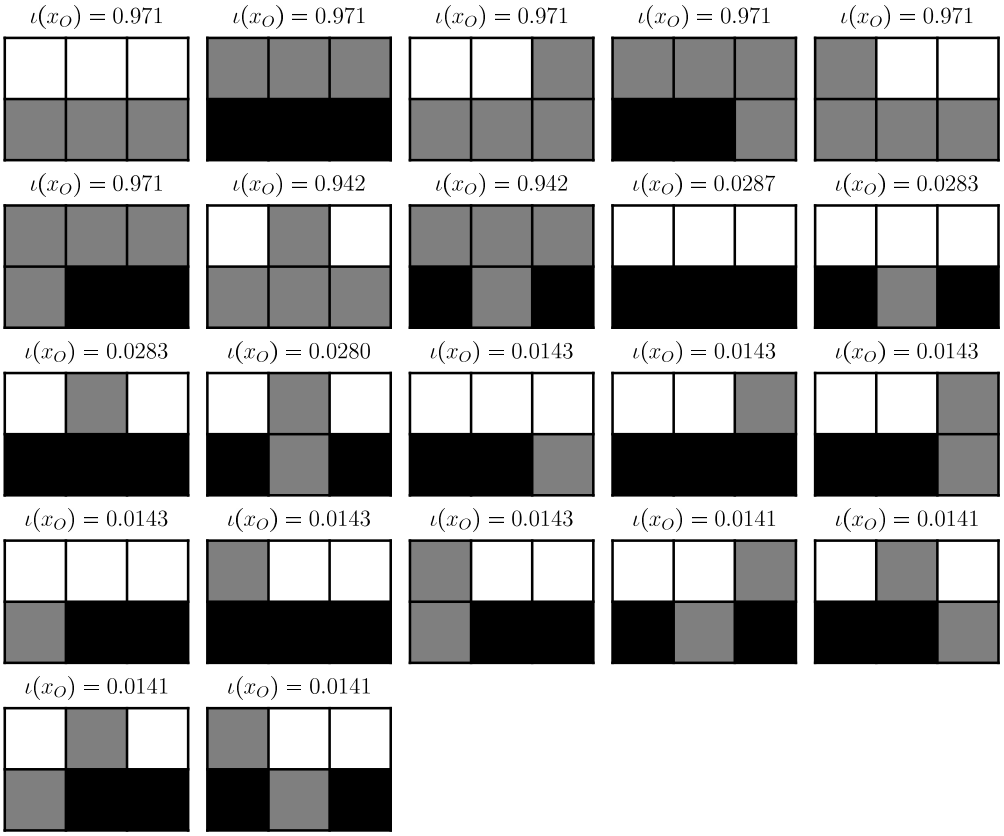


Figure 18. All distinct completely integrated composite patterns on the first trajectory x_V^1 of MC^{ϵ}. The value of complete local integration is indicated above each pattern. See Figure 12 for colouring conventions.

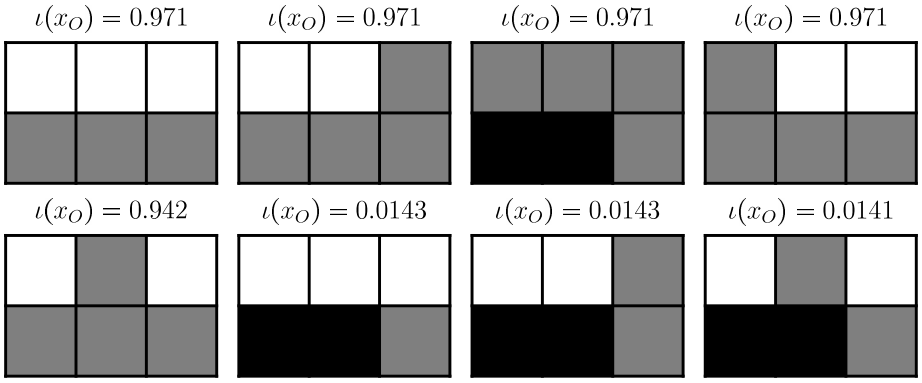


Figure 19. All distinct completely integrated composite patterns on the second trajectory x_V^2 of MC^{ϵ}. The value of complete local integration is indicated above each pattern.

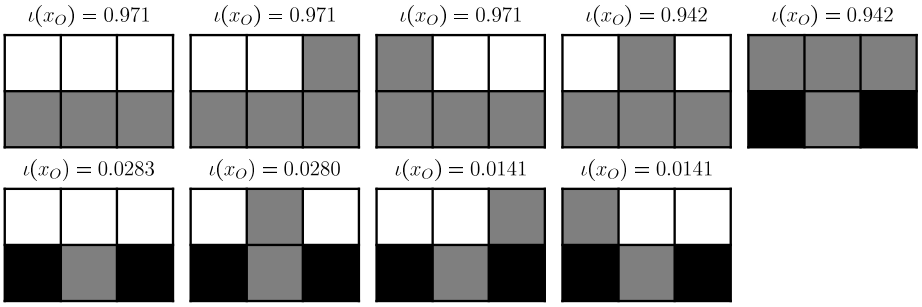


Figure 20. All distinct completely integrated composite patterns on the third trajectory x_V^3 of MC^ϵ. The value of complete local integration is indicated above each pattern.

5. Discussion

In Section 3.1 we have argued for the use of patterns as candidates for entities. Patterns can be composed of arbitrary spatially and temporally extended parts of trajectories. We have seen in Theorem 1 that they are distinct from arbitrary subsets of trajectories. The important insight here is that patterns are structures that occur within trajectories but this cannot be said of sets of trajectories.

One of the main target applications of patterns is in time-unrolled Bayesian networks of cellular automata like those in Figure 1. Patterns in such Bayesian networks become spatiotemporal patterns like those used to describe the glider, block, and blinker in the Game of Life cellular automaton by Beer [14]. We would also like to investigate whether the latter spatiotemporal patterns are ι-entities. However, at the present state of the computational models and, without approximations, this was out of reach computationally. We will discuss this further below.

In Section 3.3 we defined SLI and in Section 3.3 gave its expression for deterministic Bayesian networks (including cellular automata) as well. We also established the least upper bound of SLI with respect to a partition π of cardinality n for a pattern x_A with probability q. This upper bound is achieved if each of the blocks x_b in the partition π occur if and only if the whole pattern x_O occurs. This is compatible with our interpretation of entities since in this case clearly the occurrence of any part of the pattern leads necessarily to the occurrence of the entire pattern (and not only vice versa).

We also presented a candidate for a greatest lower bound of SLI with respect to a partition of cardinality n for a pattern with probability q. Whether this is the greatest lower bound or not it shows a case for which SLI is always negative. This happens if either the whole pattern x_A occurs (with probability q) or one of the "almost equal" patterns occurs, each with identical probability. A pattern y_A is almost equal to x_A with respect to π in this sense if it only differs at one of the blocks $b \in \pi$ i.e., if $y_A = (x_{A\setminus b}, z_b)$ where $z_b \neq x_b$. This construction makes as many parts as possible (i.e., all but one) occur as many times as possible without the whole pattern occurring. This creates large marginalised probabilities $p_b(x_b)$ for each part/block which means that their product probability also becomes large.

Beyond these quantitative interpretations an interpretation of the greatest lower bound candidate seems difficult. A more intuitive candidate for the opposite of an integrated pattern seem to be patterns with independent parts. i.e., zero SLI but quantitatively these are not on the opposite end of the SLI spectrum. A more satisfying interpretation of the presented candidate is still to be found.

We also proved the disintegration theorem which relates states that the refinement-free partitions of a trajectory among those partitions achieving a particular SLI value consist of ι-entities only, where an ι-entity is a pattern with positive CLI. This theorem allows us to interpret the ι-entities in new ways and may lead to a more formal or quantitative justification of ι-entities. It is already a first step in this direction since it establishes a special role of the ι-entities within trajectories of Bayesian networks. A further justification would tell us what in turn the refinement-free partitions can be used for. We have discussed a possible direction for further investigation in detail in Section 3.5. This tried to connect the ι-entities with a coding problem.

In Section 4 we investigated SLI and CLI in three simple example sets of random variables. We found that if the random variables are all independently distributed the according entities are just all the possible $x_j \in \mathcal{X}_j$ of each of the random variables $X_j \in \{X_i\}_{i \in V}$. This is what we would expect from an entity criterion. There are no entities with any further extent than a single random variable and each value corresponds to a different entity.

For the simple Markov chain $MC^{=}$ composed out of two independent and constant processes we presented the entire disintegration hierarchy and the Hasse diagrams of each disintegration level ordered by refinement. The Hasse diagrams reflected the highly symmetric dynamics of the Markov chain via multiple identical components. For the refinement-free disintegration hierarchy we then get multiple partitions at the same disintegration level as well. Different partitions of the trajectory imply overlapping blocks which in the case of the refinement-free partition are ι-entities. So in general the ι-entities at a particular disintegration level are not necessarily unique and can overlap. We also saw in Figure 11 that the same ι-entities can occur on multiple disintegration levels.

The ι-entities of $MC^{=}$ included the expected three timestep constant patterns within each of the two independent processes. It also included the two timestep parts of these constant patterns. This may be less expected. It shows that parts of ι-entities can be ι-entities themselves. We note that these "sub-entities (those that are parts of larger entities) are always on a different disintegration level than their" super-entities (the larger entities). We can speculate that the existence of such sub- and super-entities on different disintegration levels may find an interpretation through multicellular organisms or similar structures. However, the overly simplistic examples here only serve as basic models for the potential phenomena, but are still far too simplistic to warrant any concrete interpretation in this direction.

We also looked at a version of $MC^{=}$ perturbed by noise, denoted MC^{ϵ}. We found that the entities of $MC^{=}$ remain the most strongly integrated entities in MC^{ϵ}. At the same time new entities occur. So we observe that in MC^{ϵ} the entities vary from one trajectory to another (Figures 18–20). We also observe spatially extended entities i.e., entities that extend across both (formerly independent) processes. We also observe entities that switch from one process to the other (from top row to bottom row or vice versa). The capacity of entities to exhibit this behaviour may be necessary to capture the movement or metabolism of entities in more realistic scenarios. In Biehl et al. [8] we argued that these properties are important and showed that they hold for a crude approximation of CLI (namely for SLI with respect to $\pi = \mathbf{0}$) but not for the full CLI measure.

We established that the ι-entities:

- correspond to fixed single random variables for a set of independent random variables,
- can vary from one trajectory to another,
- and can change the degrees of freedom that they occupy over time,
- can be ambiguous at a fixed level of disintegration due to symmetries of the system,
- can overlap at the same level of disintegration due to this ambiguity,
- can overlap across multiple levels of disintegration i.e., parts of ι-entities can be ι-entities again.

In general the examples we investigated concretely are too small to sufficiently support the concept of positive CLI as an entity criterion. Due to the extreme computational burden, this may remain the case for a while. For a straightforward calculation of the minimum SLI of a trajectory of a Bayesian network $\{X_i\}_{i \in V}$ with $|V| = k$ nodes we have to calculate the SLI with respect to $\mathcal{B}_k$ partitions. According to (Bruijn [43], p. 108) the Bell numbers $\mathcal{B}_n$ grow super-exponentially. Furthermore, to evaluate the SLI we need the joint probability distribution of the Bayesian network $\{X_i\}_{i \in V}$. Naively, 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 gives us around 10^{30} petabytes of storage needed for this probability

distribution. We are sceptical that the exact evaluation of reasonably large systems can be achieved even with non-naive methods. This suggests that formal proofs may be the more promising way to investigate SLI and CLI further.

Acknowledgments: Part of the research was performed during Martin Biehl's time as a JSPS International Research Fellow and as an ELSI Origins Network (EON) long term visitor. Daniel Polani was supported in part by the EC Horizon 2020 H2020-641321 socSMCs FET Proactive project.

Author Contributions: Martin Biehl, Takashi Ikegami, and Daniel Polani conceived the problem and the measures, and wrote the paper; Martin Biehl proved the theorems and conceived and calculated the examples. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SLI Specific local integration
CLI Complete local integration

Appendix A. Kronecker Delta

The Kronecker–delta is used in this paper to represent deterministic conditional distributions.

Definition A1 (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*

$$\delta_x(C) := \begin{cases} 1 & \text{if } x \in C, \\ 0 & \text{else.} \end{cases} \tag{A1}$$

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

$$\delta_x(\bar{x}) := \begin{cases} 1 & \text{if } x \in \{\bar{x}\}, \\ 0 & \text{else.} \end{cases} \tag{A2}$$

$$= \begin{cases} 1 & \text{if } x = \bar{x}, \\ 0 & \text{else.} \end{cases} \tag{A3}$$

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

$$p(y|x) = \delta_{f(x)}(y), \tag{A4}$$

then

$$p(y) = \sum_x p_Y(y|x) p_X(x) \tag{A5}$$

$$= \sum_x \delta_{f(x)}(y) p_X(x) \tag{A6}$$

$$= \sum_x \delta_x(f^{-1}(y)) p_X(x) \tag{A7}$$

$$= \sum_{x \in f^{-1}(y)} p_X(x) \tag{A8}$$

$$= p_X(f^{-1}(y)). \tag{A9}$$

Appendix B. Refinement and Partition Lattice Examples

This appendix recalls the definitions of set partitions, refinement and coarsening of set partitions, and Hasse diagrams. It also shows the Hasse diagram of an example partition lattices. The definitions are due to Grätzer [28].

Definition A2. *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* $x_1 \cap x_2 = \emptyset$,
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 A3. *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 A4 (Refinement and coarsening)**.** *We define the binary relation* $\trianglelefteq$ *between partitions* $\pi, \rho \in \mathfrak{L}(\mathcal{X})$ *as:*

$$\pi \trianglelefteq \rho \text{ if } x_1 \equiv_\pi x_2 \text{ implies } x_1 \equiv_\rho x_2. \tag{A10}$$

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

Remarks:

- 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.
- Examples are contained in the Hasse diagrams (defined below) shown in Figure A1.

Definition A5 (Hasse diagram)**.** *A Hasse diagram is a visualisation of a poset (including lattices). Given a poset* A *ordered by* $\lhd$ *the Hasse diagram represents the elements of* A *by dots. The dots representing the elements are arranged in such a way that if* $a, b \in A$, $a \neq b$, *and* $a \lhd b$ *then the dot representing* a *is drawn below the dot representing* b. *An edge is drawn between two elements* $a, b \in A$ *if* $a \lhd\!\!:\, b$, *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 $a \lhd b$ but not $a \lhd\!\!:\, b$.
- 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.
- For an example Hasse diagrams see Figure A1.

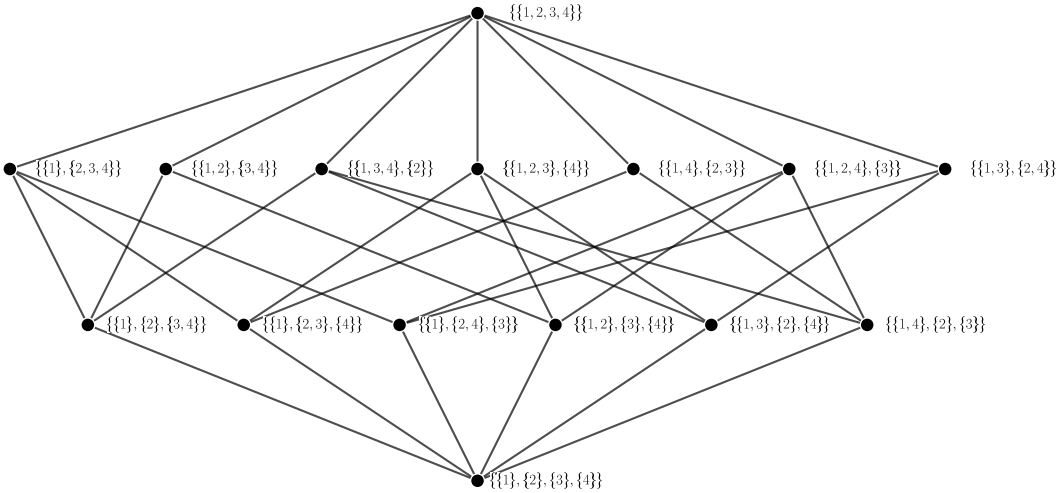


Figure A1. Hasse diagrams of the partition lattice of the four element set.

Appendix C. Bayesian Networks

Intuitively a Bayesian network is a graph representation of the inter-dependencies of a set of random variables. Recall that any joint probability distribution over a set $\{X_i\}_{i\in I}$ with $I = \{1, ..., n\}$ of random variables can always be written as a product of conditional probabilities:

$$p_I(x_1, ..., x_n) = \prod_{i=1}^{n-1} p_i(x_i|x_{i+1}, ..., x_n)p(x_n). \tag{A11}$$

This also holds for any reordering of the indices $i \mapsto f(i)$ with $f : \{1, ...n\} \to \{1, ...n\}$ bijective.

In many cases however this factorisation can be simplified. Often some of the conditional probabilities $p(x_i|x_{i+1}, ..., x_n)$ do not depend on all variables $\{x_{i+1}, ..., x_n\}$ listed in the product of Equation (A11). For example, X_1 might only depend on X_2 so we would have $p(x_1|x_2, ..., x_n) = p(x_1|x_2)$. Note that the latter conditional probability is determined by fixing $|\mathcal{X}_2|(|\mathcal{X}_1| - 1)$ probabilities whereas the former needs $\prod_{i=1}^{n} |\mathcal{X}_{i+1}|(|\mathcal{X}_1| - 1)$ probabilities to be fixed. This means the number of parameters (the probabilities) of the joint distribution $p(x_1, ..., x_n)$ is often much smaller than suggested by Equation (A11). One way to encode this simplification and make sure that we are dealing only with joint probabilities that reflect the dependencies we allow are Bayesian networks. These can be defined as follows. First we define graphs that are factorisation compatible with joint probability distributions over a set of random variables and then define the Bayesian networks as pairs of joint probability distributions and a factorisation compatible graph.

Definition A6. *A directed acyclic graph $G = (V, E)$ with nodes V and edges E is* factorisation compatible *with the joint probabilities the probabilities of a probability distribution $p_V : \mathcal{X}_V \to [0, 1]$ iff the latter can be factorised in the way suggested by G which means:*

$$p_V(x_V) = \prod_{i\in V} p(x_i|x_{\mathrm{pa}(i)}). \tag{A12}$$

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

Remark:

- In general there are multiple directed acyclic graphs that are factorisation compatible with the same probability distribution. For example, 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 Equation (A12) becomes Equation (A11) which always holds.

Definition A7 (Bayesian network). *A* Bayesian network *is a (here assumed 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 factorisation 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.*

Remarks:

- On top of constituting the vertices of the graph G the set V is also assumed to be totally ordered in an (arbitrarily) fixed way. Whenever we use a subset $A \subset V$ to index a sequence of variables in the Bayesian network (e.g., in $p_A(x_A)$) we order A according to this total order as well.
- Since $\{X_i\}_{i\in V}$ is finite and G is acyclic there is a set V_0 of nodes without parents.

Definition A8 (Mechanism). *Given a Bayesian network $\{X_i\}_{i\in V}$ with index set V for each node with parents i.e., for each node $i \in V \setminus V_0$ (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)*

$$\sum_{x_i \in \mathcal{X}_i} p_i(x_i|x_{\mathrm{pa}(i)}) = 1. \tag{A13}$$

Remarks:

- 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 Equation (A12). Note that in Equation (A12) the nodes in V_0 are not explicit as they are just factors $p_i(x_i|x_{\mathrm{pa}(i)})$ with $\mathrm{pa}(i) = \emptyset$.
- To construct a Bayesian network, take graph $G = (V, E)$ and equip each node $i \in (V \setminus V_0)$ 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 Equation (A12):

$$p_V(x_V) = \prod_{i\in V\setminus V_0} p_i(x_i|x_{\mathrm{pa}(i)}) \prod_{j\in V_0} p_j(x_j). \tag{A14}$$

Appendix C.1. Deterministic Bayesian Networks

Definition A9 (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*

$$p_i(x_i|x_{\mathrm{pa}(i)}) = \delta_{f_i(x_{\mathrm{pa}(i)})}(x_i) = \begin{cases} 1 & \text{if } x_i = f_i(x_{\mathrm{pa}(i)}), \\ 0 & \text{else.} \end{cases} \tag{A15}$$

Definition A10 (Deterministic Bayesian network). *A Bayesian network $\{X_i\}_{i\in V}$ is deterministic if all its mechanisms are deterministic.*

Theorem A1. *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 Equation (A12), the definition of conditional probabilities, and using the deterministic mechanisms we have:

$$p_{V\setminus V_0}(x_{V\setminus V_0}|x_{V_0}) = \prod_{i\in V\setminus V_0} p_i(x_i|x_{\mathrm{pa}(i)}) \tag{A16}$$

$$= \prod_{i\in V\setminus V_0} \delta_{f_i(x_{\mathrm{pa}(i)})}(x_i). \tag{A17}$$

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 A2 (Pattern probability in a deterministic Bayesian network)**.** *Given a deterministic Bayesian network (Definition A10) and uniform initial distribution $p_{V_0} : \mathcal{X}_{V_0} \to [0,1]$ the probability of the occurrence of a pattern x_A is:*

$$p_A(x_A) = \frac{N(x_A)}{|\mathcal{X}_{V_0}|} \tag{A18}$$

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 A1) 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 $A \cap V_0 = A_0$, $A \setminus V_0 = A_r$, $B := V \setminus A$, $B \cap V_0 = B_0$, and $B \setminus V_0 = B_r$. Then

$$p_A(x_A) = \sum_{\bar{x}_B} p_V(x_A, \bar{x}_B) \tag{A19}$$

$$= \sum_{\bar{x}_{B_0}, \bar{x}_{B_r}} p_V(x_{A_r}, \bar{x}_{B_r}|x_{A_0}, \bar{x}_{B_0}) p_{V_0}(x_{A_0}, \bar{x}_{B_0}) \tag{A20}$$

$$= \sum_{\bar{x}_{B_0}, \bar{x}_{B_r}} \delta_{f_{V\setminus V_0}(x_{A_0}, \bar{x}_{B_0})}(x_{A_r}, \bar{x}_{B_r}) p_{V_0}(x_{A_0}, \bar{x}_{B_0}) \tag{A21}$$

$$= \sum_{\bar{x}_{B_r}} \sum_{\{\bar{x}_{B_0} : (x_{A_0}, \bar{x}_{B_0}) \in f^{-1}_{V\setminus V_0}(x_{A_r}, \bar{x}_{B_r})\}} p_{V_0}(x_{A_0}, \bar{x}_{B_0}) \tag{A22}$$

$$= \frac{1}{|\mathcal{X}_{V_0}|} \sum_{\bar{x}_{B_r}} |\{\bar{x}_{B_0} \in \mathcal{X}_{B_0} : (x_{A_0}, \bar{x}_{B_0}) \in f^{-1}_{V\setminus V_0}(x_{A_r}, \bar{x}_{B_r})\}| \tag{A23}$$

$$= \frac{1}{|\mathcal{X}_{V_0}|} N(x_A) \tag{A24}$$

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 a 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 a 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 $1/|\mathcal{X}_{V_0}|$.

Appendix C.2. Proof of Theorem 2

Proof. Follows by replacing the probabilities $p_O(x_O)$ and $p_b(x_b)$ in Equation (22) with their deterministic expressions from (Theorem A2), i.e., $p_A(x_A) = N(X_A)/|\mathcal{X}_{V_0}|$. Then:

$$\begin{aligned}
\mathrm{mi}_\pi(x_O) :&= \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)} && \text{(A25)}\\
&= \log \frac{\frac{N(x_O)}{|\mathcal{X}_{V_0}|}}{\prod_{b\in\pi} \frac{N(x_b)}{|\mathcal{X}_{V_0}|}} && \text{(A26)}\\
&= \log \frac{\frac{N(x_O)}{|\mathcal{X}_{V_0}|}}{|\mathcal{X}_{V_0}|^{-|\pi|}\prod_{b\in\pi} N(x_b)} && \text{(A27)}\\
&= \log \frac{|\mathcal{X}_{V_0}|^{|\pi|-1}N(x_O)}{\prod_{b\in\pi} N(x_b)} && \text{(A28)}\\
&= (|\pi|-1)\log|\mathcal{X}_{V_0}| - \log\frac{\prod_{b\in\pi} N(x_b)}{N(x_O)}. && \text{(A29)}
\end{aligned}$$

□

Appendix D. Proof of Theorem 1

Proof. Given a set of random variables $\{X_i\}_{i\in V}$, a subset $\mathcal{D} \subseteq \mathcal{X}_V$ cannot be represented by a pattern of $\{X_i\}_{i\in V}$ if and only if there exists $A \subseteq V$ with $\mathcal{D}_A \subset \mathcal{X}_A$ (proper subset) and $|\mathcal{D}_A| > 1$, i.e., if neither all patterns at A are possible nor a unique pattern at A is specified by $\mathcal{D}$.

We first show that if there exists $A \subseteq V$ with $\mathcal{D}_A \subset \mathcal{X}_A$ and $|\mathcal{D}_A| > 1$ then there is no pattern $\tilde{x}_B \in \bigcup_{C\subseteq V} \mathcal{X}_C$ with $\mathcal{D} = \mathcal{T}(\tilde{x}_B)$. Then we show that if no such A exists then there is such a pattern $\tilde{x}_B$.

Since $\mathcal{D}_A > 1$ we have $x_A, \bar{x}_A \in \mathcal{D}_A \subset \mathcal{X}_A$ with $x_A \neq \bar{x}_A$. Next note that we can write any pattern $\tilde{x}_B$ as

$$\tilde{x}_B = (\tilde{x}_{B\setminus A}, \tilde{x}_{B\cap A}). \tag{A30}$$

If $B \cap A \neq \emptyset$ we can see since $\tilde{x}_{B\cap A}$ must take a single value it cannot contain $\mathcal{D}$ since there are trajectories in $\mathcal{D}$ taking value $x_{B\cap A}$ on $B \cap A$ and trajectories in $\mathcal{D}$ taking values $\bar{x}_{B\cap A}$. More formally, we must have either $\tilde{x}_{B\cap A} = x_A$ or $\tilde{x}_{B\cap A} \neq x_A$. First, let $\tilde{x}_{B\cap A} = x_A$ but then $\mathcal{T}(\bar{x}_A) \nsubseteq \mathcal{T}(\tilde{x}_B)$ so $\mathcal{D} \nsubseteq \mathcal{T}(\tilde{x}_B)$. Next choose $\tilde{x}_{B\cap A} \neq x_A$ but then $\mathcal{T}(x_A) \nsubseteq \mathcal{T}(\tilde{x}_B)$ so also $\mathcal{D} \nsubseteq \mathcal{T}(\tilde{x}_B)$. So we must have $B \cap A = \emptyset$.

Now we show that if $B \cap A = \emptyset$ there are trajectories in $\tilde{x}_B$ that are not in $\mathcal{D}$. We construct one explicitly by fixing its value on A to the value in $\mathcal{X}_A$ that is not in $\mathcal{D}_A$ and its value on B to $\tilde{x}_B$. More formally: choose $y_A \in \mathcal{X}_A \setminus \mathcal{D}_A$, then $y_A \neq x_A$ and $y_A \neq \bar{x}_A$. This is always possible since $\mathcal{D}_A \subset \mathcal{X}_A$ (proper subset). Then consider a trajectory $\hat{x}_V = (\tilde{x}_B, y_A, \check{x}_D)$ with arbitrary $\check{x}_D \in \mathcal{X}_D$ where $D = V \setminus (B \cup A)$. Then $\hat{x}_V \in \mathcal{T}(\tilde{x}_B)$ but $\hat{x}_V \notin \mathcal{D}$.

Conversely, we show how to construct $\tilde{x}_B$ if no such A exists. the idea is just to fix all random variables where $|\mathcal{D}_A| = 1$ and leave them unspecified where $D_A = \mathcal{X}_A$. More formally: if there exists no $A \subseteq V$ with $\mathcal{D}_A \subset \mathcal{X}_A$ and $|\mathcal{D}_A| > 1$, then for each $C \subseteq V$ either $\mathcal{D}_C = \mathcal{X}_C$ or $|\mathcal{D}_C| = 1$. Then let $B = \bigcup\{C \subseteq V : |\mathcal{D}_C| = 1\}$ then $|\mathcal{D}_B| = 1$ so that we can define $\tilde{x}_B$ as the unique element in $\mathcal{D}_B$. Then if $y_V \in \mathcal{D}$ we have $y_B = \tilde{x}_B$ so $\mathcal{D} \subseteq \mathcal{T}(\tilde{x}_B)$. If $z_V \in \mathcal{T}(\tilde{x}_B)$ we have $z_B = \tilde{x}_B \in \mathcal{D}_B$ and for $A \subseteq V$ with $A \cap B = \emptyset$ by construction of B we have $\mathcal{D}_A = \mathcal{X}_A$ such that $\mathcal{D}_{V\setminus B} = \mathcal{X}_{V\setminus B}$ which means $z_{V\setminus B} \in \mathcal{D}_{V\setminus B}$ and therefore $z_V \in \mathcal{D}$ and $\mathcal{T}(\tilde{x}_B) \subseteq \mathcal{D}$. So this gives $\mathcal{T}(\tilde{x}_B) = \mathcal{D}$. □

References

1. Gallois, A. Identity over Time. In *The Stanford Encyclopedia of Philosophy*; Zalta, E.N., Ed.; Metaphysics Research Laboratory, Stanford University: Stanford, CA, USA, 2012.
2. Grand, S. *Creation: Life and How to Make It*; Harvard University Press: Harvard, MA, USA, 2003.
3. Pascal, R.; Pross, A. Stability and its manifestation in the chemical and biological worlds. *Chem. Commun.* **2015**, *51*, 16160–16165.
4. Orseau, L.; Ring, M. Space-Time Embedded Intelligence. In *Artificial General Intelligence*; Number 7716 in Lecture Notes in Computer Science; Bach, J., Goertzel, B., Iklé, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 209–218.
5. Barandiaran, X.E.; Paolo, E.D.; Rohde, M. Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-temporality in Action. *Adapt. Behav.* **2009**, *17*, 367–386.
6. Legg, S.; Hutter, M. Universal Intelligence: A Definition of Machine Intelligence. *arXiv* **2007**, arXiv: 0712.3329.
7. Boccara, N.; Nasser, J.; Roger, M. Particlelike structures and their interactions in spatiotemporal patterns generated by one-dimensional deterministic cellular-automaton rules. *Phys. Rev. A* **1991**, *44*, 866–875.
8. Biehl, M.; Ikegami, T.; Polani, D. Towards information based spatiotemporal patterns as a foundation for agent representation in dynamical systems. In Proceedings of the Artificial Life Conference, Cancun, Mexico, 2016; The MIT Press: Cambridge, MA, USA, 2016; pp. 722–729.
9. McGill, W.J. Multivariate information transmission. *Psychometrika* **1954**, *19*, 97–116.
10. Amari, S.I. Information geometry on hierarchy of probability distributions. *IEEE Trans. Inf. Theory* **2001**, *47*, 1701–1711.
11. Lizier, J.T. *The Local Information Dynamics of Distributed Computation in Complex Systems*; Springer: Berlin/Heidelberg: Germany, 2012.
12. Tononi, G.; Sporns, O. Measuring information integration. *BMC Neurosci.* **2003**, *4*, 31.
13. Balduzzi, D.; Tononi, G. Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework. *PLoS Comput. Biol.* **2008**, *4*, e1000091.
14. Beer, R.D. Characterizing autopoiesis in the game of life. *Artif. Life* **2014**, *21*, 1–19.
15. Fontana, W.; Buss, L.W. "The arrival of the fittest": Toward a theory of biological organization. *Bull. Math. Biol.* **1994**, *56*, 1–64.
16. Krakauer, D.; Bertschinger, N.; Olbrich, E.; Ay, N.; Flack, J.C. The Information Theory of Individuality. *arXiv* **2014**, arXiv:1412.2447.
17. Bertschinger, N.; Olbrich, E.; Ay, N.; Jost, J. Autonomy: An information theoretic perspective. *Biosystems* **2008**, *91*, 331–345.
18. Shalizi, C.R.; Haslinger, R.; Rouquier, J.B.; Klinkner, K.L.; Moore, C. Automatic filters for the detection of coherent structure in spatiotemporal systems. *Phys. Rev. E* **2006**, *73*, 036104.
19. Wolfram, S. Computation theory of cellular automata. *Commun. Math. Phys.* **1984**, *96*, 15–57.
20. Grassberger, P. Chaos and diffusion in deterministic cellular automata. *Phys. D Nonlinear Phenom.* **1984**, *10*, 52–58.
21. Hanson, J.E.; Crutchfield, J.P. The attractor—Basin portrait of a cellular automaton. *J. Stat. Phys.* **1992**, *66*, 1415–1462.
22. Pivato, M. Defect particle kinematics in one-dimensional cellular automata. *Theor. Comput. Sci.* **2007**, *377*, 205–228.
23. Lizier, J.T.; Prokopenko, M.; Zomaya, A.Y. Local information transfer as a spatiotemporal filter for complex systems. *Phys. Rev. E* **2008**, *77*, 026110.
24. Flecker, B.; Alford, W.; Beggs, J.M.; Williams, P.L.; Beer, R.D. Partial information decomposition as a spatiotemporal filter. *Chaos Interdiscip. J. Nonlinear Sci.* **2011**, *21*, 037104.
25. Friston, K. Life as we know it. *J. R. Soc. Interface* **2013**, *10*, 20130475.
26. Balduzzi, D. Detecting emergent processes in cellular automata with excess information. *arXiv* **2011**, arXiv:1105.0158.
27. Hoel, E.P.; Albantakis, L.; Marshall, W.; Tononi, G. Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neurosci. Conscious.* **2016**, *2016*, niw012.
28. Grätzer, G. *Lattice Theory: Foundation*; Springer: New York, NY, USA, 2011.

29. Ceccherini-Silberstein, T.; Coornaert, M. Cellular Automata and Groups. In *Encyclopedia of Complexity and Systems Science*; Meyers, R.A., Ed.; Springer: New York, NY, USA, 2009; pp. 778–791.
30. Busic, A.; Mairesse, J.; Marcovici, I. Probabilistic cellular automata, invariant measures, and perfect sampling. *arXiv* **2010**, arXiv: 1010.3133.
31. Beer, R.D. The cognitive domain of a glider in the game of life. *Artif. Life* **2014**, *20*, 183–206.
32. Beer, R.R. *Autopoiesis and Enaction in the Game of Life*; The MIT Press: Cambridge, MA, USA, 2016; p. 13.
33. Noonan, H.; Curtis, B. Identity. In *The Stanford Encyclopedia of Philosophy*; Zalta, E.N., Ed.; Metaphysics Research Laboratory, Stanford University: Stanford, CA, USA, 2014.
34. Hawley, K. Temporal Parts. In *The Stanford Encyclopedia of Philosophy*; Zalta, E.N., Ed.; Metaphysics Research Laboratory, Stanford University: Stanford, CA, USA, 2015.
35. Ay, N. Information Geometry on Complexity and Stochastic Interaction. *Entropy* **2015**, *17*, 2432–2458.
36. MacKay, D.J. *Information Theory, Inference and Learning Algorithms*; Cambridge University Press: Cambridge, UK, 2003.
37. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley: Hoboken, NJ, USA, 2006.
38. Tononi, G. An information integration theory of consciousness. *BMC Neurosc.* **2004**, *5*, 42.
39. Von Eitzen, H. Prove $(1-(1-q)/n)^n \geq q$ for $0 < q < 1$ and $n \geq 2$ a Natural Number. Mathematics Stack Exchange, 2016. Available online: http://math.stackexchange.com/q/1974262 (accessed on 18 October 2016).
40. Bullen, P.S. *Handbook of Means and Their Inequalities*; Springer Science+Business Media: Dordrecht, The Netherlands, 2003.
41. Kolchinsky, A.; Rocha, L.M. Prediction and modularity in dynamical systems. In *Advances in Artificial Life, ECAL*; The MIT Press: Cambridge, MA, USA, 2011; pp. 423–430.
42. Pemmaraju, S.; Skiena, S. *Computational Discrete Mathematics: Combinatorics and Graph Theory with Mathematica®*; Cambridge University Press: Cambridge, UK, 2009.
43. De Bruijn, N.G. *Asymptotic Methods in Analysis*; Dover Publications: New York, NY, USA, 2010.