Article

Specific and Complete Local Integration of Patterns in Bayesian Networks

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

[1] Araya, 2F Mori 15 Building, 2-8-10 Toranomon, Minato-ku, Tokyo 105-0001, Japan
[2] Department of Computer Science, University of Hertfordshire, Hatfield AL10 9AB, United Kingdom
[3] Department of General Systems Studies, University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan
* Correspondence: martin@araya.org

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

Contents

1. Introduction

This paper presents 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[1] 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[2]) 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 [cmp. 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 agents 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

---

[1] The authors thank Eric Smith for pointing out the example of a soap bubble.
[2] In the following, if not stated otherwise, we always mean *spatiotemporal* entities when we refer to entities.

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.

The purpose of this contribution is to establish some formal properties of the authors' previously published criterion for entities [7] in (finite[3]) Bayesian networks. While the previous work was more conceptual, the present one represents a first formal investigation. We introduce basic notions like *patterns* more rigorously and clearly distinguish them from subsets of trajectories. Then we recall our criterion for *ι-entities* i.e positive *complete local integration* (CLI) and the measure that it is constructed from i.e. *specific local integration* (SLI). In order to get a better feel for the behaviour of SLI we constructively prove the least upper bound and give a candidate for a lower bound of SLI (and thereby CLI). The main contribution, however, is the *disintegration theorem*. This theorem relates the SLI of an entire space-time trajectory to the complete local integration of parts of this trajectory. This provides a new perspective on the ι-entities fulfilling the CLI criterion, revealing them as distinct structures within trajectories of dynamics induced by Bayesian networks and related systems like cellular automata and multivariate Markov chains. This is a significant step forward since previous motivations of the CLI criterion were based exclusively on intuition. The disintegration theorem is based on the newly defined disintegration hierarchy and its refinement-free version. The disintegration hierarchy orders partitions of entire trajectories according to their SLI values. Its refinement-free discards the partitions of the disintegration hierarchy that have a refining partition with a lower or equal SLI. According to the disintegration theorem blocks of partitions in the refinement-free disintegration hierarchy are ι-entities. Finally, we show the disintegration hierarchies and resulting ι-entities for simple example systems. These provide the first "sanity checks" for the CLI criterion. In the previous work [7] only a crude approximation of CLI was computed.

Apart from the study of formal concepts of identity, the disintegration theorem might is also of interest for researchers studying multi-information [8,9]. Since the latter is just the expectation value of specific local information.

As of now, there are few attempts at definitions of spatiotemporally extended entities in Bayesian networks or related systems like cellular automata. However, formally our measure is very closely related to others in the literature. Furthermore, there are publication which can be interpreted as alternative definitions of spatiotemporally extended entities.

From a formal perspective our measures are a combination of existing concepts. We start with an arbitrary set of random variables (in our case usually a time-unrolled Bayesian network). First note that multi-information can be slightly generalised from splitting a set of random variables into all its singleton subsets to splitting it into an arbitrary partition π [as done e.g. for the similar stochastic interaction in 10]. Then we can apply the method of localising information theoretical measures described by Lizier [11] to such a π-multi-information. This results in specific local integration (SLI), called specific because of the partition dependence and local because of the localisation of the multi-information. Finally, combine this with the weakest-link approach of finding the partition such that some measure of integration takes its minimum. This has been proposed for bipartitions [12] and arbitrary partitions [13] in the context of integrated information theory [14]. The result of this

[3] We only consider finite systems in this paper.

combination is the *complete* local integration (CLI). To our knowledge CLI has been proposed for the first time by us in [7]. The latter publication does not include any of the formal or numerical results in the present paper.

Conceptually, our work is most closely related to Beer [15]. The notion of spatiotemporal patterns used there to capture blocks, blinkers, and gliders is equivalent to the *patterns* we define more formally here. In this work there is also a non-formally stated criterion for identity over time and entities.

The construction of the entities proceeds roughly as follows. First the maps from the Moore neighbourhood to the next state of a cell are classified into five classes of *local processes*. Then these are used to reveal the dynamical structure in the transitions from one time-slice (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 reoccurence of the organisation.

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

We note that using the entity criterion we propose does not need similar choices to be made since it is not based on the reocurrence 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.

It is worth noting that viewing entities/objects/individuals as occurring within a trajectory is in contrast to an approach that models them as stochastic processes interacting with an "environment". In the stochastic process view the individual is represented by a sequence of random variables. At each individual timestep the corresponding random variable represents the state of the the individual at that timestep. Every sequence of states of these random variables corresponds to one of the possible time evolutions of the individual. From the point of view in this paper individuals are represented not by sequences of random variables but by sequences of *states* of random variables. This means that in our view another sequence of states of the same sequence of random variables is not another possible time evolution of the same individual/entity. In fact in our case it may not even correspond to the time evolution of an individual/entity at all. In many cases the stochastic process representation of entities deals with entities for which the problem of identity over time is solved. This is the case if the sequence of random variables that represents the entity is defined a priori. For example if we have two possibly interacting stochastic processes $\{X_t, Y_t\}_{t\in T}$ where at least one of the processes represents an entity. A notable exception is Krakauer *et al.* [16]. Based on information theoretic notions of autonomy and closure due to Bertschinger *et al.* [17] Krakauer *et al.* present criteria to detect which random variables in a multivariate random process should be seen as representing parts of the individual and which parts of the environment. The principle of autonomy proposed by Bertschinger *et al.* [17] 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.

At the most basic level the intuition behind entities is that some spatiotemporal patterns are different from others. This also underlies spatiotemporal filtering. We here discuss work that is similar on this most basic level and then indicate how ι-entities essentially differ due to their explicit treatment of identity over time.

Defining (and usually finding) more important spatiotemporal patterns or structures (also called coherent structures) has a long history in the theory of cellular automata and distributed dynamical systems. As Shalizi *et al.* [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.

Applying any one of the definitions (or associated methods) proposed by [18,23,24] to the time-evolution (what we call a *trajectory*) of a cellular automaton assigns each cell or group of cells[4] $A_{(j,t)}$ indexed by j at each time t a value that measures a property of the cells $A_{(j,t)}$. The result is then a "filtered" time evolution of the cells (or groups of cells) in the cellular automaton where each cell j at time t now takes its value of the measured property. These filtered time evolutions then highlight according spatiotemporally extended structures. These are often gliders and domains. However, these methods make no claim about the identity over time of the revealed patterns. This means that there is no criterion given that tells us which cells and their values at time t_1 and which cells and their values at time t_2 are part of the same entity or object. For isolated gliders this may not seem like a problem but whenever gliders collide it is not clear whether they both loose their identity and become a new thing (or no thing) or whether one of them survived the collision and maybe just changed direction. Spatiotemporal filtering provides no answers to these questions. Our approach can provide these answers by returning not only sequences of spatial structures but spatiotemporal entities which explicitly determine identity over time. Whether the answers are the "right ones" and in which sense remains an open question however.

We note here that the approach of Friston [25] using Markov blankets of the non-time-unrolled dynamics has the same shortcoming as the spatiotemporal filters. For each timestep it returns a partition of all degrees of freedom [5] into internal, sensory, active, and external degrees. However, it does not provide a way to resolve identity over time. Since only a single Markov blanket is defined in Friston [25] ambiguities due to multiple colliding Markov blankets are ruled out. This limitation to single Markov can only be temporary if the approach is ever to describe more than a single organism. At that point ambiguous situations will occur and won't be resolved by the Markov blanket only.

Among the research related to integrated information theory (IIT) there are some approaches that lead to spatiotemporal coarse-grainings of multivariate systems like the ones we are also interested in. The resulting "grains" are spatially and temporally extended and so could be interpreted as spatiotemporal entities. The goal of these approaches, however, is not to define entities. They are more directly aimed at establishing the optimal spatiotemporal scale to describe the dynamics of a system. The earliest such work seems to be Balduzzi [26]. More recently Hoel *et al.* [2728] proposed approaches with similar aims but somewhat different formalism. While Hoel *et al.* [27] produces a trajectory independent coarse-graining and therefore cannot describe entities that occupy one set of random variables in one trajectory and another (possibly partly intersecting) set of random variables in another trajectory, Balduzzi [26] and Hoel *et al.* [28] lead to trajectory dependent coarse-grainings. Compared to our approach it is notable that the spatiotemporal grains are determined by their

---

[4] The group of cells $A_{(j,t)}$ might be the past light-cone of the cell at (j,t) as in the case of local statistical complexity [18]. This makes no difference to the following argument however as the result is still just a (discrete) value at (j,t).

[5] In our formalism the degrees of freedom of the dynamical system in Friston [25] correspond to the spatial random variables.

interactions with other grains. In our case the entities are determined first and foremost by their internal relations. One of the main consequences of this is that the occurrence of an entity does not only depend on the occurrence of the entity's pattern itself but also on the spatiotemporal context in which this pattern occurs. In other words 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. This is related to the philosophical discussion about identity across possible worlds [29].

Some 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.

We also note that it is uncertain whether formal definitions of entities are needed, for example every combination of events could be an entity (this is what is called "unrestricted mereological composition" in philosophy) so no further criterion for entities is needed. The only thing required for a formalisation of unrestricted mereological composition is a formal definition of all events and all their combinations. In the present work this would correspond to the set of all patterns.

In conclusion, this paper further characterises the measures of specific and complete local integration. In particular it reveals a special role of ι-entities within trajectories via the disintegration hierarchy. More precisely, ι-entities are blocks in the finest partitions of a trajectory that achieve a particular value of SLI. Since this allows further interpretations of ι-entities (see Section 5) we hope this can lead to an entirely formal justification of the CLI criterion.

2. Notation and background

In this section we briefly introduce our notation for sets of random variables[6] 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 $p_A = p_i$,*

[6] Since every set of random variables can be seen as a Bayesian network and vice versa we use these terms interchangeably.

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.*
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*[7] ρ*.*
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 background on partition lattices see e.g. Grätzer [30]. 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. 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 patterns are different from the subsets of trajectories of a set of random variables.

---

[7] This 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 $\pi \preceq \xi \preceq \rho$.

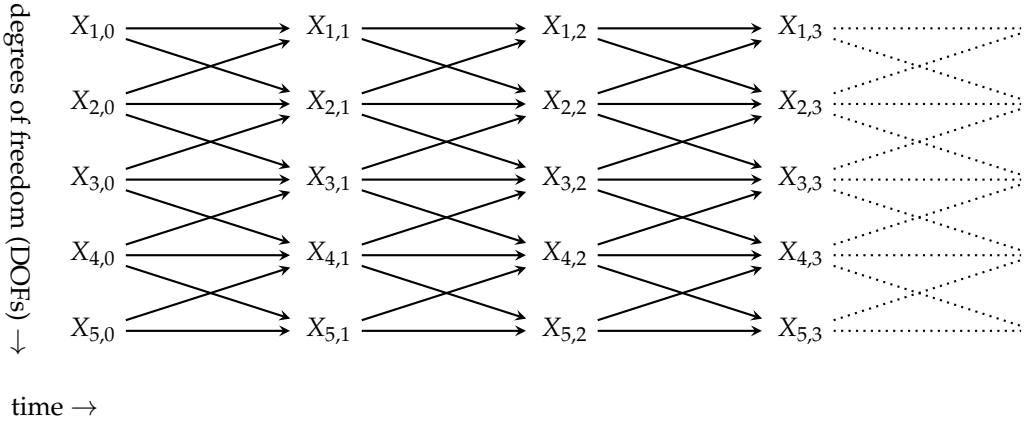


Figure 1. *First time steps of a Bayesian network representing a multivariate dynamical system (or multivariate Markov chain) $\{X_i\}_{i\in V}$. Here we used $V = J \times T$ with J indicating spatial degrees of freedom and T the temporal extension. Then each node is indexed by a tuple (j,t) as shown. The shown edges are just an example, any two nodes within the same or subsequent columns can be connected as long as the target of the edge is not on the left of its origin.*

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 then used to prove the main result, the *disintegration theorem*. This relates the SLI of a 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 Fig. 1 for an example of a Bayesian network of such a system). 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 a Game of Life cellular automaton. The time evolutions over multiple steps of the cells attributed to a glider or a block [see 15, for a principled way to attribute cells to these] should be patterns but also arbitrary choices of subsets of cells (and their values) possibly extending over multiple timesteps together.

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 (and trajectories) formally in the following way.

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. *Patterns $X_V = x_V$ at V which fix the complete set $\{X_i\}_{i\in V}$ of random variables are called* trajectories.
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 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.

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 Ceccherini-Silberstein and Coornaert [31] and to "cylinders" as defined in Busic *et al.* [32]. However the notions there are explicitly restricted to single time-slices of (probabilistic) cellular automata. Our notion of patterns purposely lifts this restriction. Our notion is inspired by the usage of the term spatiotemporal pattern in Beer [153334] and we suggest that it formalises this notion.

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 equal. Formally:

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

While patterns define subsets of $\mathcal{X}_V$ not every subset of $\mathcal{X}_V$ is defined (or captured) by a pattern.

To see this we next show how to construct a subset of $\mathcal{X}_V$ that is not a pattern. First 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*

$$\mathcal{D}_A := \{x_A \in \mathcal{X}_A : \exists \bar{x}_V \in \mathcal{X}_V, \bar{x}_A = x_A\}. \tag{10}$$

With this we 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 $\mathcal{D} \in \mathcal{X}_V$ 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 C. □

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

Peer-reviewed version available at Entropy 2017, 19, , 230; doi:10.3390/e19050230

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[8]) 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 thesis is referred to as four-dimensionalism or perdurantism in philosophical discussions [see e.g. 35]. 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:

$$\min_{b \subset O} \frac{p_{O \setminus b}(x_{O \setminus b} | x_b)}{p_{O \setminus b}(x_{O \setminus b})} > 1. \tag{11}$$

This is equivalent to

$$\min_{b \subset O} \frac{p_O(x_O)}{p_{O \setminus b}(x_{O \setminus b}) p_b(x_b)} > 1. \tag{12}$$

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

$$\min_{\pi \in \mathfrak{L}_2(O)} \frac{p_O(x_O)}{\prod_{b \in \pi} p_b(x_b)} > 1. \tag{13}$$

[8] The two components of the problem of identity are also called *synchronic* and *diachronic* identity [29].

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

$$p_O(x_O) =: p_{b_1,b_2}(x_{b_1}, x_{b_2}) = p_{b_1}(x_{b_1})p_{b_2}(x_{b_2}). \tag{14}$$

Which would give us

$$\frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)} = 1 \tag{15}$$

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

$$p_O(x_O) = \prod_{c\in\xi} p_c(x_c) \tag{16}$$

seems to suggest that the pattern x_O is not an entity but instead composite of three parts. We can therefore generalise Eq. (13) 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

$$\min_{\pi\in\mathfrak{L}(O)\setminus\mathbf{1}_O} \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)} > 1. \tag{17}$$

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.

Anticipating the definition of CLI in Definition 8 we then define the ι-entities.

Definition 5 (ι-entity)**.** *Given a multivariate Markov chain $\{X_i\}_{i\in V}$ a pattern x_O is a ι-entity if*

$$\min_{\pi\in\mathfrak{L}(O)\setminus\mathbf{1}_O} \mathrm{mi}_\pi(x_O) > 0. \tag{18}$$

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

Definition 6 (*ci*-entity-set)**.** *Given a multivariate Markov chain $\{X_i\}_{i\in V}$ the ι-entity-set is the entity-set*

$$\mathfrak{E}_\iota(\{X_i\}_{i\in V}) := \{x_O \in \bigcup_{A\subseteq V} \mathcal{X}_A : \iota(x_O) > 0\}. \tag{19}$$

In the next sections we look at more formal properties of expressions that occur in these definitions.

3.3. Specific local integration

This section introduces the specific local integration (SLI). It also proves its upper bounds constructively and constructs an example of negative SLI.

3.3.1. General and deterministic case

Definition 7 (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*

$$\mathrm{mi}_\pi(x_O) := \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)}. \tag{20}$$

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

Theorem 2 (Deterministic specific local integration)**.** *Given a deterministic Bayesian network (Definition B.5), 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*

$$\mathrm{mi}_\pi(x_O) = (|\pi| - 1) \log |\mathcal{X}_{V_0}| - \log \frac{\prod_{b \in \pi} N(x_b)}{N(x_O)}. \tag{21}$$

Proof. See Appendix B.2. □

3.3.2. Upper bounds

In this section we present the upper bounds of SLI. These are of general interest, constructive proofs which may serve to familiarise the reader with the measure of SLI are given in Appendix D.

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. In the subsequent theorem we show that this property in general gives the upper bound of SLI if the cardinality of the partition is fixed. This leads directly tho the upper bound if the cardinality of the partition is not fixed in the next theorem. Finally we give the expressions of the bounds in the deterministic case for convenient reference.

Theorem 3 (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*

$$\max_{\{\{X_i\}_{i \in V} : \exists x_O, p_O(x_O) = q\}} \max_{\{\pi : |\pi| = n\}} \mathrm{mi}_\pi(x_O) \leq -(n-1) \log q. \tag{22}$$

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

$$p_b(x_b) = p_O(x_O) = q. \tag{23}$$

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

Proof. See Appendix D.1 □

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 $|\pi| \leq |O|$.

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).

Theorem 4. *For any given probability $q < 1$ and cardinality $|\pi| = n > 1$ of a partition π there exists a pattern x_O in a Bayesian network $\{X_i\}_{i\in V}$ such that $q = p_O(x_O)$ and*

$$\mathrm{mi}_\pi(x_O) = \log \frac{q}{\left(1 - \frac{1-q}{n}\right)^n} < 0. \tag{24}$$

Proof. See Appendix D.2 □

Remarks:

- The achieved value in Eq. (24) 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.
- An interpretation of the construction is that patterns which either occur as a whole or (with uniform probability) missing exactly one part always have negative SLI.

3.4. Complete local integration

We now define complete local integration (CLI) which forms the criterion distinguishing arbitrary patterns from *ι-entities*. SLI is defined with respect to a particular partition and a pattern x_A may have high SLI with respect to one partition but negative SLI with respect to another (see Section 4). Therefore SLI seems to be ill suited to define entities since then the property of being an entity would depend on how we decide to split it into parts. In order to deal with this problem we follow Tononi and Sporns [12], Balduzzi and Tononi [13], Tononi [36] who have encountered similar problems with another measure of integration. The idea is to find the partition with respect to which the pattern is least integrated and use the SLI with respect to this as the value of CLI. In other words the CLI is the SLI of x_O with respect to the partition that *disintegrates* x_O the most.

Definition 8 ((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 $\pi \in \mathfrak{L}(O) \setminus \mathbf{1}_O$:*

$$\iota(x_O) := \min_{\pi\in\mathfrak{L}(O)\setminus\mathbf{1}_O} \mathrm{mi}_\pi(x_O). \tag{25}$$

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* [37] 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 [37] and references therein. An analysis of which approach is best suited for the local integration measure presented here is beyond the scope of this paper.
- 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. Also, the candidate for the greatest lower bound of SLI is also a candidate for the greatest lower bound of CLI.

3.5. 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 the special role of patterns with positive CLI for the SLI of an entire trajectory of the system. For our purposes this theorem allows further interpretations of the measure of CLI which will be discussed in Section 5.

Definition 9 (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. $$\mathfrak{D}_1(x_V) := \operatorname*{arg\,min}_{\pi\in\mathfrak{L}(V)} \mathrm{mi}_\pi(x_V) \tag{26}$$
2. *and for $i > 1$:*
$$\mathfrak{D}_i(x_V) := \operatorname*{arg\,min}_{\pi\in\mathfrak{L}(V)\setminus\mathfrak{D}_{\prec i}(x_V)} \mathrm{mi}_\pi(x_V). \tag{27}$$

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 10. *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*

$$\mathfrak{E}_{\lhd\pi} := \{\xi \in \mathfrak{E} : \xi \lhd \pi\}. \tag{28}$$

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

Definition 11 (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. $$\mathfrak{D}_1^{\blacktriangleleft}(x_V) := \{\pi \in \mathfrak{D}_1(x_V) : \mathfrak{D}_1(x_V)_{\lhd\pi} = \emptyset\}, \tag{29}$$
2. *and for $i > 1$:*
$$\mathfrak{D}_i^{\blacktriangleleft}(x_V) := \{\pi \in \mathfrak{D}_i(x_V) : \mathfrak{D}_{\prec i}(x_V)_{\lhd\pi} = \emptyset\} \tag{30}$$

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 5 (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 $|\xi| > 1$. But then

$$\mathrm{mi}_\pi(x_V) = \log \frac{p_V(x_V)}{p_b(x_b)\prod_{c\in\pi\setminus b} p_c(x_c)} \tag{31}$$

$$\geq \log \frac{p_V(x_V)}{\prod_{d\in\xi} p_d(x_d)\prod_{c\in\pi\setminus b} p_c(x_c)} \tag{32}$$

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

$$\mathrm{mi}_\pi(x_V) = \log \frac{p_V(x_V)}{\prod_{d\in\xi} p_d(x_d)\prod_{c\in\pi\setminus b} p_c(x_c)}. \tag{33}$$

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

$$\mathrm{mi}_\pi(x_V) > \log \frac{p_V(x_V)}{\prod_{d\in\xi} p_d(x_d)\prod_{c\in\pi\setminus b} p_c(x_c)}. \tag{34}$$

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

$$\mathrm{mi}_\rho(x_V) < \mathrm{mi}_\pi(x_V), \tag{35}$$

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

ad 2 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

$$\mathrm{mi}_{\xi|_A}(x_A) = \log \frac{p_A(x_A)}{\prod_{c\in\xi|_A} p_c(x_c)} > 0 \tag{36}$$

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

$$\begin{aligned} \mathrm{mi}_{\xi}(x_V) &= \log \frac{p_V(x_V)}{\prod_{c\in\xi|_A} p_c(x_c) \prod_{i\in V\setminus A} p_i(x_i)} && (37)\\ &> \log \frac{p_V(x_V)}{p_A(x_A) \prod_{i\in V\setminus A} p_i(x_i)} && (38)\\ &= \mathrm{mi}_{\pi^A}(x_V). && (39) \end{aligned}$$

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 $\pi^A \in \mathfrak{D}^{\blacktriangleleft}_{i^A}(x_V)$.

□

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 5) 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 $\pi \in \mathfrak{L}(V)$:

$$\mathrm{mi}_{\pi}(x_V) = 0. \tag{40}$$

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 $\mathfrak{D}^{\blacktriangleleft}_1 = \{\mathbf{0}\}$. 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

- $$\mathrm{pa}(j,t) = \begin{cases} \emptyset & \text{if } t = 0, \\ \{(j,t-1)\} & \text{else,} \end{cases} \tag{41}$$
- $$p_{j,t}(x_{j,t}|x_{j,t-1}) = \delta_{x_{j,t-1}}(x_{j,t}) = \begin{cases} 1 & \text{if } x_{j,t} = x_{j,t-1}, \\ 0 & \text{else,} \end{cases} \tag{42}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 March 2017 doi:10.20944/preprints201703.0171.v1

Peer-reviewed version available at Entropy 2017, 19, , 230; doi:10.3390/e19050230

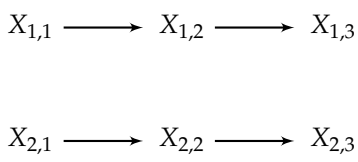
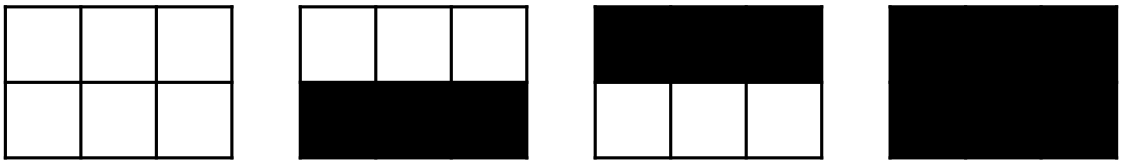


Figure 2. *Bayesian network of $MC^{=}$. There is no interaction between the two processes.*



(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 3. *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.*

- $$p_{j,0}(x_{j,0}) = 1/4. \tag{43}$$

The Bayesian network can be seen in Fig. 2.

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:

$$x_V = (x_{1,0}, x_{2,0}, x_{1,1}, x_{2,1}, x_{1,2}, x_{2,2}) = \begin{cases} (0,0,0,0,0,0) & \text{if } x_{1,0} = 0, x_{2,0} = 0; \\ (0,1,0,1,0,1) & \text{if } x_{1,0} = 0, x_{2,0} = 1; \\ (1,0,1,0,1,0) & \text{if } x_{1,0} = 1, x_{2,0} = 0; \\ (1,1,1,1,1,1) & \text{if } x_{1,0} = 1, x_{2,0} = 1. \end{cases} \tag{44}$$

Each of these trajectories has probability $p_V(x_V) = 1/4$ and all other trajectories have $p_V(x_V) = 0$. We call the four trajectories the *possible trajectories*. We visualise the possible trajectories as a grid with each cell corresponding to one variable. The spatial indices are constant across rows and time-slices V_t correspond to the columns. A white cell indicates a 0 and a black cell indicates a 1. This results in the grids of Fig. 3.

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 March 2017 doi:10.20944/preprints201703.0171.v1

Peer-reviewed version available at Entropy 2017, 19, , 230; doi:10.3390/e19050230

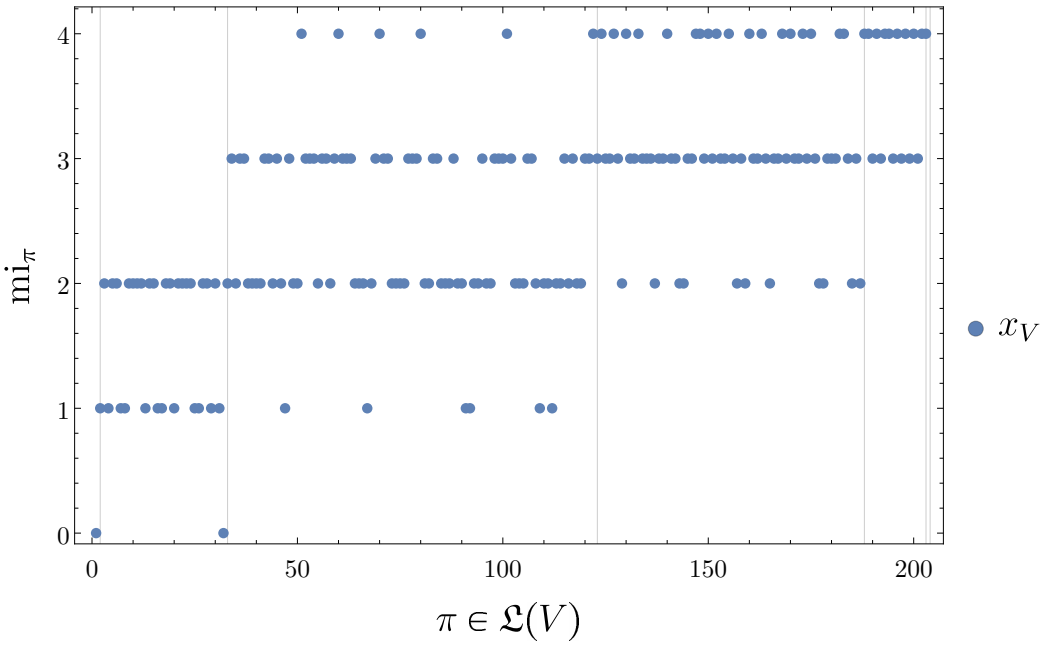


Figure 4. *Specific local integrations* $\mathrm{mi}_\pi(x_V)$ *of any of the four trajectories* x_V *seen in Fig. 3 with respect to all* $\pi \in \mathfrak{L}(V)$*. The partitions are ordered according to an enumeration with increasing cardinality* $|\pi|$ *(see 38, chap. 4.3.3 for the method). We indicate with vertical lines at what partitions the cardinality* $|\pi|$ *increases by one.*

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

(45)

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 7:

$$\mathrm{mi}_\pi(x_V) := \log \frac{p_V(x_V)}{\prod_{b\in\pi} p_b(x_b)}. \tag{46}$$

In the case of $MC^=$ the SLI values with respect to each partition do not depend on the trajectories. For an overview we plotted the values of SLI with respect to each partition $\pi \in \mathfrak{L}(V)$ for any trajectory of $MC^=$ in Fig. 4. We can see in Fig. 4 that the cardinality does not determine the value of SLI. At the same time there seems to be a trend to higher values of SLI with increasing cardinality of the partition. We can also observe that only five different values of SLI are attained by partitions on this trajectory. We will collect these classes of partitions with equal SLI values in the disintegration hierarchy next.

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 9) and look at the Hasse diagrams of each of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 March 2017 doi:10.20944/preprints201703.0171.v1

Peer-reviewed version available at Entropy 2017, 19, , 230; doi:10.3390/e19050230

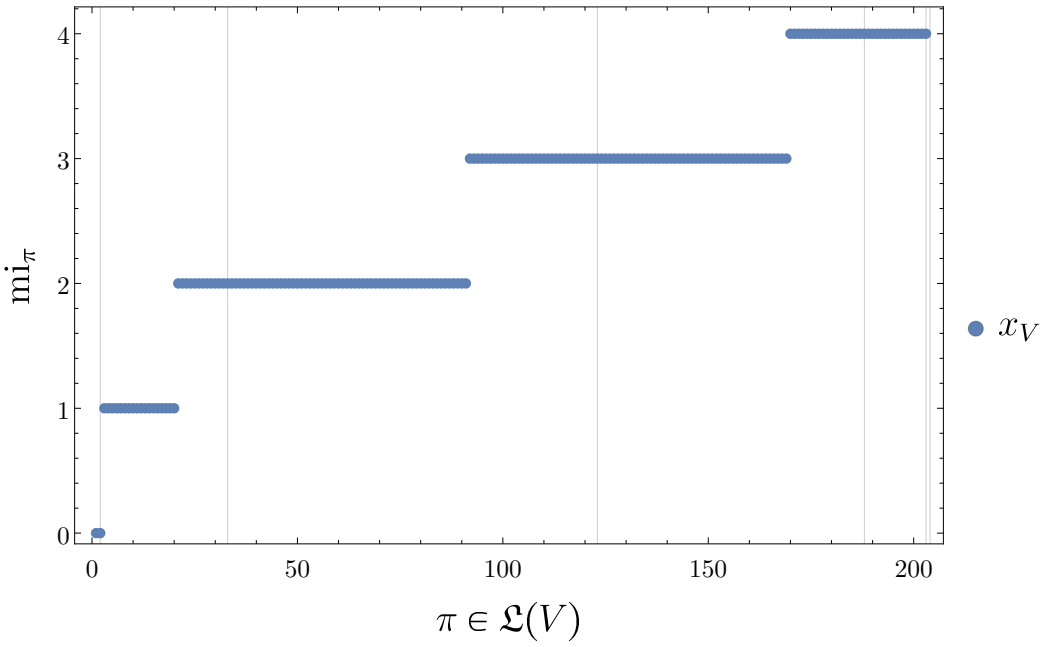


Figure 5. *Same as Fig. 4 but with the partitions sorted according to increasing SLI.*

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 Fig. 5. 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:

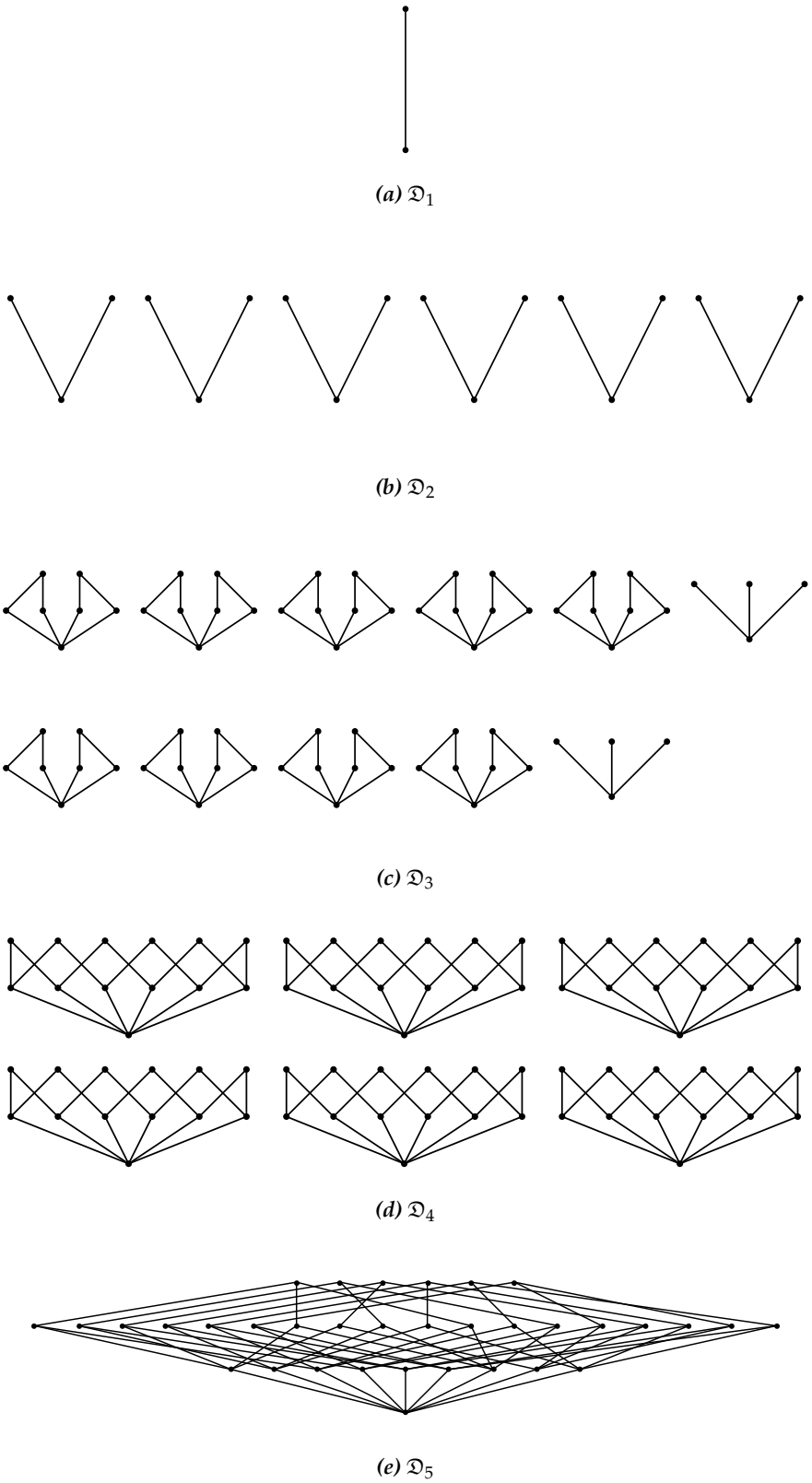
i	1	2	3	4	5
mi_π	0	1	2	3	4
$\lvert\mathfrak{D}_i\rvert$	2	18	71	78	34

(47)

Next we look at the Hasse diagram of each of those disintegration levels (partially ordered by refinement). Since the disintegration levels are subsets of the partition lattice $\mathfrak{L}(V)$, they are in general not lattices by themselves. The Hasse diagrams visualise the set of partitions in each disintegration level partially ordered by refinement $\lhd$. Recall that in Hasse diagrams of such posets the partitions are arranged such that if $\pi \neq \xi$ and $\pi \lhd \xi$ then π is drawn below ξ. Also, an edge is drawn from partition π to ξ if one *covers* the other e.g. if $\pi \lhd\!: \xi$ (note the colon indicating the covering relation). The Hasse diagrams are shown in Fig. 6. We see immediately that within each disintegration level apart from the first and the last the Hasse diagrams contain multiple connected components.

Furthermore, within a disintegration level the connected components often have the same Hasse diagrams. For example in $\mathfrak{D}_2$ (Fig. 6b) 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 Fig. 7.

Recall that due to the disintegration theorem (Theorem 5) we are interested especially in partitions that do not have refinements at their own or any preceding (i.e. lower indexed) disintegration level. These partitions consist of blocks that are completely integrated i.e. all possible partitions of each of the blocks results in a positive SLI value or is a single node of the Bayesian network. The refinement-free disintegration hierarchy $\mathfrak{D}^{\blacktriangleleft}(x_V)$ contains only these partitions and is shown in a Hasse diagram in Fig. 8.



(a) $\mathfrak{D}_1$

(b) $\mathfrak{D}_2$

(c) $\mathfrak{D}_3$

(d) $\mathfrak{D}_4$

(e) $\mathfrak{D}_5$

Figure 6. *Hasse diagrams of the five disintegration levels of the trajectories of* $MC^{=}$*. Every vertex corresponds to a partition and edges indicate that the lower partition refines the higher one.*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 March 2017 doi:10.20944/preprints201703.0171.v1

Peer-reviewed version available at Entropy 2017, 19, , 230; doi:10.3390/e19050230

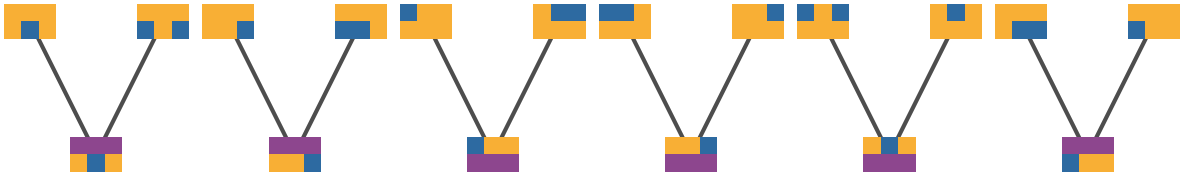


Figure 7. *Hasse diagram of $\mathfrak{D}_2$ of $MC^{=}$ trajectories. Here we visualise the partitions at each vertex. The blocks of a partition are the cells of equal colour. Note that we can obtain all six disconnected components from one by symmetry operations that are respected by the joint probability 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.*

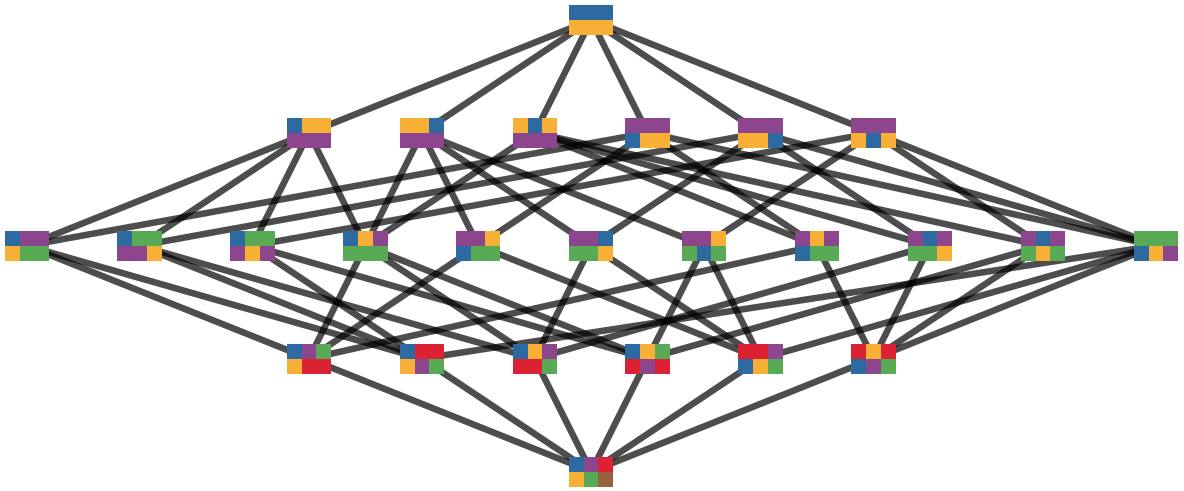


Figure 8. *Hasse diagrams of the refinement-free disintegration hierarchy $\mathfrak{D}^{\blacktriangleleft}$ of $MC^{=}$ trajectories. Here we visualise the partitions at each vertex. The blocks of a partition are the cells of equal colour. It turns out that partitions that are on the same horizontal level in this diagram correspond exactly to a level in the refinement-free disintegration hierarchy $\mathfrak{D}^{\blacktriangleleft}$. The i-th horizontal level starting from the top corresponds to $\mathfrak{D}_i^{\blacktriangleleft}$. Take for example the second horizontal level from the top. The partitions on this level are just the minimal elements of the poset $\mathfrak{D}_2$ which was visualised in Fig. 7. To connect this to Fig. 6 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 of that disintegration level.*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 March 2017 doi:10.20944/preprints201703.0171.v1

Peer-reviewed version available at Entropy 2017, 19, , 230; doi:10.3390/e19050230

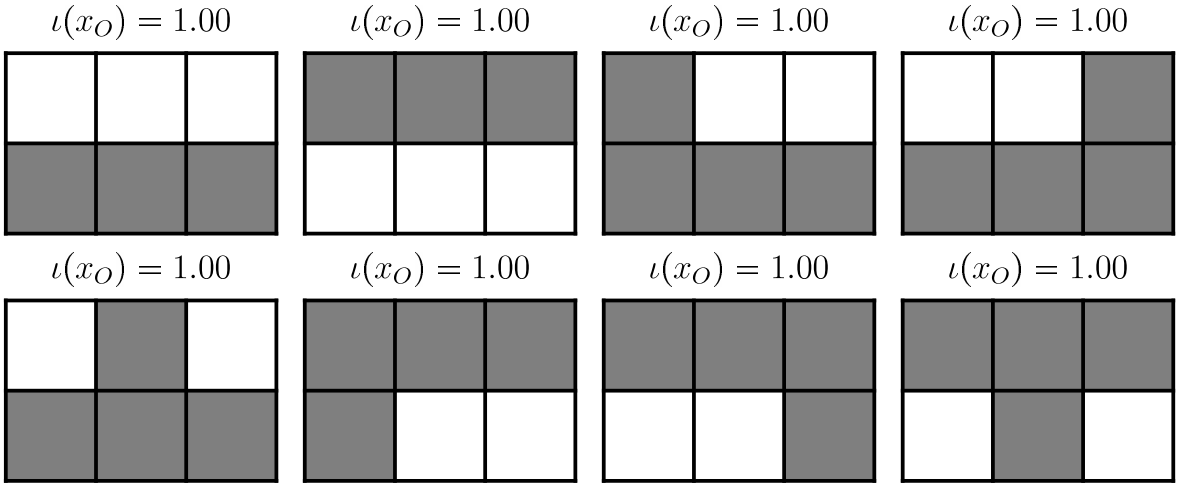


Figure 9. *All distinct completely integrated composite patterns (singletons are not shown) on the first possible trajectory of $MC^{=}$. The value of complete local integration is indicated above each pattern. We display patterns by colouring the cells corresponding to random variables that are not 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*.*

4.2.6. Completely integrated patterns

Having looked at the disintegration hierarchy we now make use of it by extracting the completely (locally[9]) integrated patterns of the four trajectories of $MC^{=}$. Recall that due to the disintegration theorem (Theorem 5) we know that all blocks in partitions that occur in the refinement-free disintegration hierarchy are either singletons or correspond to completely integrated patterns. If we look at the refinement-free disintegration hierarchy in Fig. 8 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 completely integrated patterns and their complete local integration (Definition 8) on the first trajectory in Fig. 9 and on the second trajectory in Fig. 10.

Since the disintegration hierarchies are the same for the four possible trajectories of $MC^{=}$ we get the same refinement-free partitions and therefore the same blocks containing the completely integrated patterns. This is apparent when comparing Figs. 9 and 10 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 Fig. 9 there is a corresponding pattern in Fig. 10 leaving the same cells grey.

If we are not interested in a particular trajectory we can also look at all different completely integrated patterns on any trajectory. For $MC^{=}$ these are shown in Fig. 11 We see that all completely integrated patterns x_O have the same value of complete local integration $\iota(x_O) = 1$. This can be explained using the deterministic expression for the SLI of Eq. (21) 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

[9] When it is clear from context that we are talking about complete local integration we drop "local" for the sake of readability.

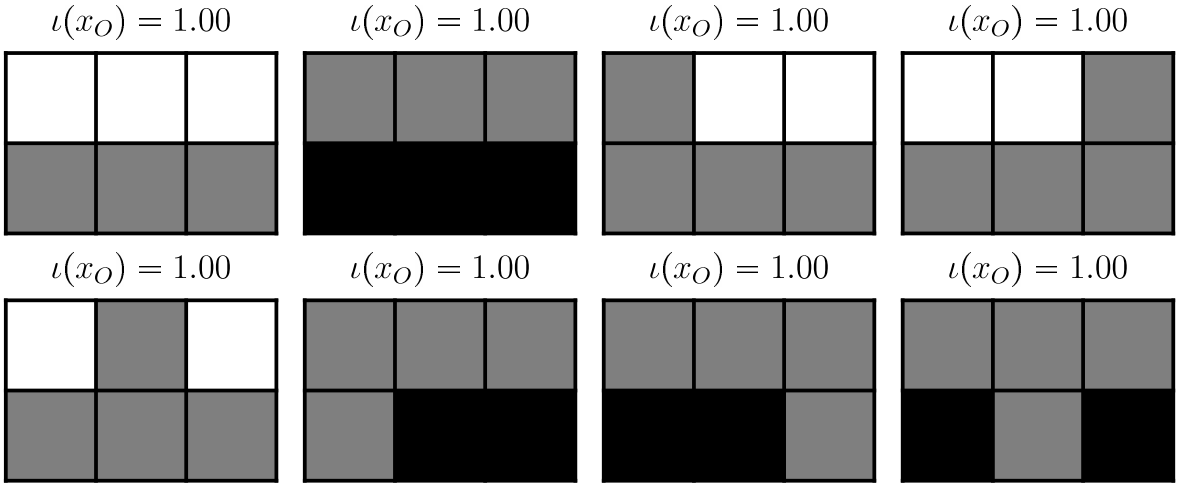


Figure 10. *All distinct completely integrated composite patterns on the second possible trajectory of $MC^{=}$. The value of complete local integration is indicated above each pattern.*

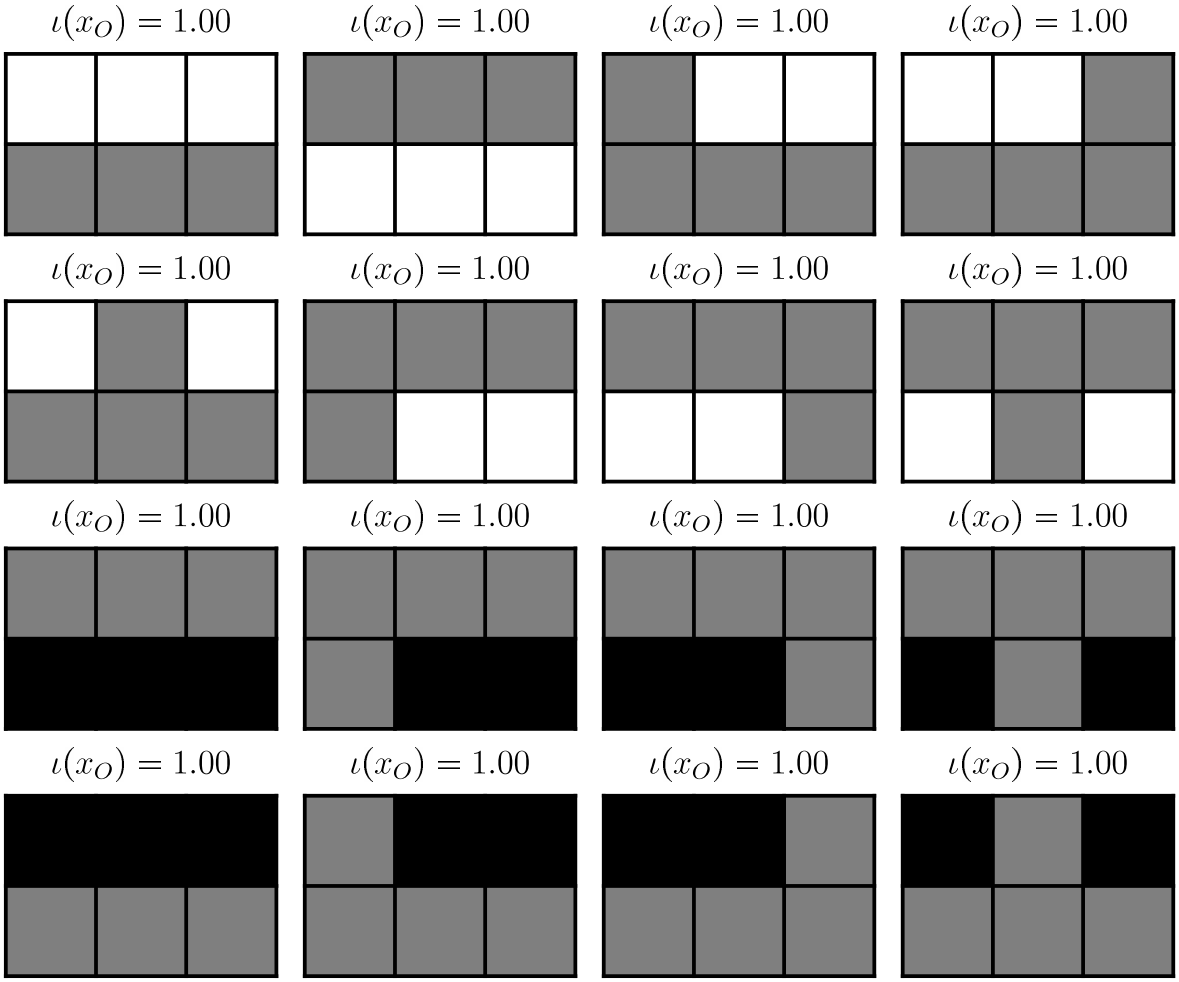


Figure 11. *All distinct completely integrated composite patterns on all four possible trajectories of $MC^{=}$. The value of complete local integration is indicated above each pattern.*

same value. This means that the number of trajectories $N(x_{j,S})$ in which any pattern $x_{j,S}$ with $S \subseteq T$ occurs is either $N(x_{j,S}) = 0$, if the pattern is impossible, or $N(x_{j,S}) = 2$ since there are two trajectories compatible with it. Note that all blocks x_b in any of the completely integrated pattern and all pattern x_O themselves are of the form $x_{j,S}$ with $S \subseteq T$. Let $N(x_{j,S}) =: N$ and plug this into Eq. (21) for an arbitrary partition π:

$$\mathrm{mi}_\pi(x_O) = (|\pi| - 1) \log |\mathcal{X}_{V_0}| - \log \frac{\prod_{b\in\pi} N(x_b)}{N(x_O)} \tag{48}$$

$$= (|\pi| - 1) \log |\mathcal{X}_{V_0}| - \log \frac{N^{|\pi|}}{N} \tag{49}$$

$$= (|\pi| - 1) \log \frac{|\mathcal{X}_{V_0}|}{N}. \tag{50}$$

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 completely integrated patterns are all limited to one of the two rows. This shows on a simple example that, as we would expect, completely integrated patterns cannot extend from one independent process to another.

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

$$P_{f(x_{1,t+1},x_{2,t+1}),f(x_{1,t},x_{2,t})} = p_{J,t+1}(x_{1,t+1}, x_{2,t+1}|x_{1,t}, x_{2,t}) \tag{51}$$

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

$$f(0,0) = 1, f(0,1) = 2, f(1,0) = 3, f(1,1) = 4. \tag{52}$$

With this convention P is

$$P = \begin{pmatrix} 1-3\epsilon & \epsilon & \epsilon & \epsilon \\ \epsilon & 1-3\epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & 1-3\epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & 1-3\epsilon \end{pmatrix} \tag{53}$$

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. In the following we set $\epsilon = 1/100$. The initial distribution is again the uniform distribution

$$p_{j,0}(x_{j,0}) = 1/4. \tag{54}$$

Writing this multivariate Markov chain as a Bayesian network is possible but the conversion is tedious. The Bayesian network one obtains can be seen in Fig. 12.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 March 2017 doi:10.20944/preprints201703.0171.v1

Peer-reviewed version available at Entropy 2017, 19, , 230; doi:10.3390/e19050230

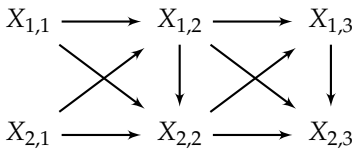
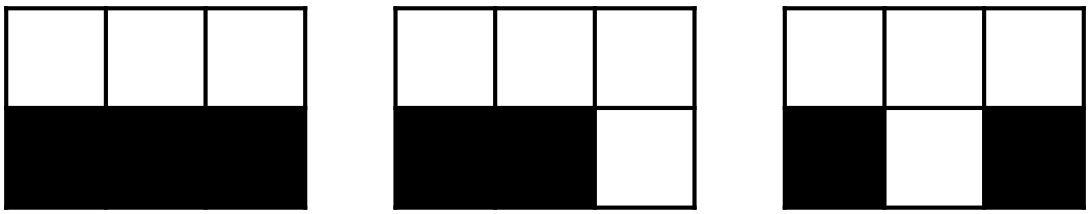


Figure 12. *Bayesian network of MC^{ϵ}.*



(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 13. *Visualisation of three trajectories of MC^{ϵ}. In each trajectory the time index increases from left to right. There are two rows corresponding to the two random variables at each time step and three columns corresponding to the three time-steps we are considering here. We can see that the first trajectory (in a) makes no ϵ-transitions, the second (in b) makes one from $t = 2$ to $t = 1$, and the third (in c) makes two.*

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 Fig. 13 and will be denoted x_V^1, x_V^2, and x_V^3 respectively. The probabilities are $p_V(x_V^1) = 0.235225, p_V(x_V^2) = 0.0024250, p_V(x_V^3) = 0.000025$.

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 Fig. 14.

It turns out that the SLI values of x_V^1 are almost the same as those of $MC^{=}$ in Fig. 4 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 completely integrated patterns 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

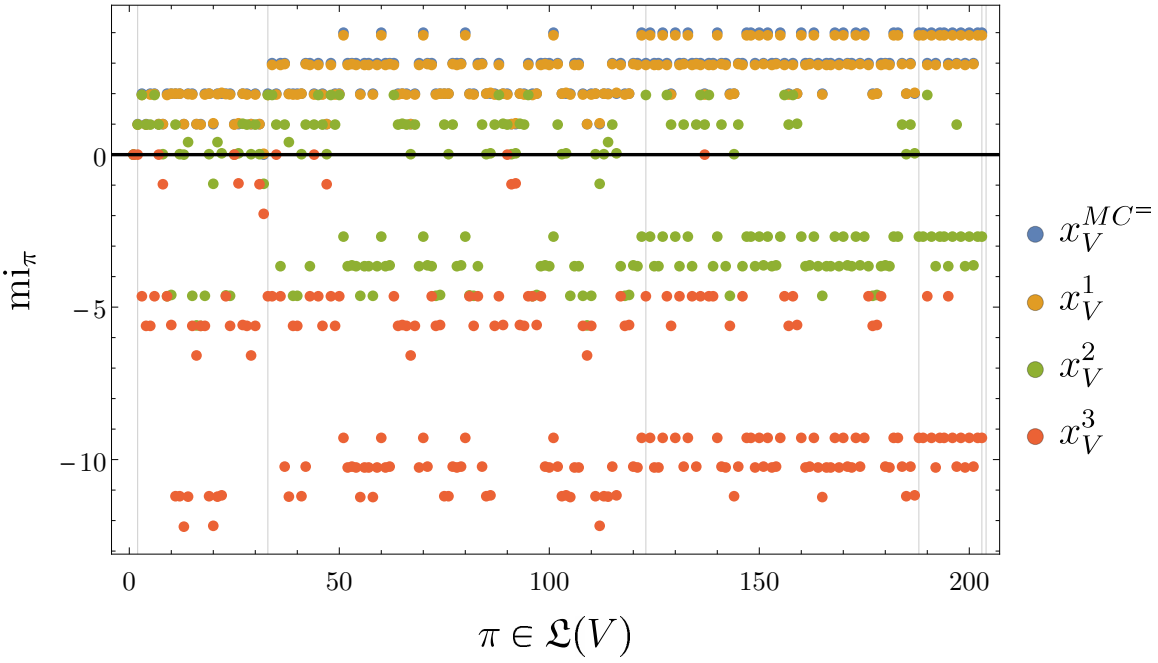


Figure 14. *Specific local integrations* $\mathrm{mi}_\pi(x_V)$ *of one of the four trajectories of* $MC^=$ *(measured w.r.t. the probability distribution of* $MC^=$*), here denoted* $x_V^{MC^=}$ *(this is the same data as in Fig. 4), and the three representative trajectories* $x_V^k, x \in \{1,2,3\}$ *of* MC^ϵ *(measured w.r.t. the probability distribution of* MC^ϵ*) seen in Fig. 13 with respect to all* $\pi \in \mathfrak{L}(V)$*. The partitions are ordered as in Fig. 4 with increasing cardinality* $|\pi|$*. Vertical lines indicate partitions where the cardinality* $|\pi|$ *increases by one. Note that the values of* $x_V^{MC^=}$ *are almost completely hidden from view by those of* x_V^1*.*

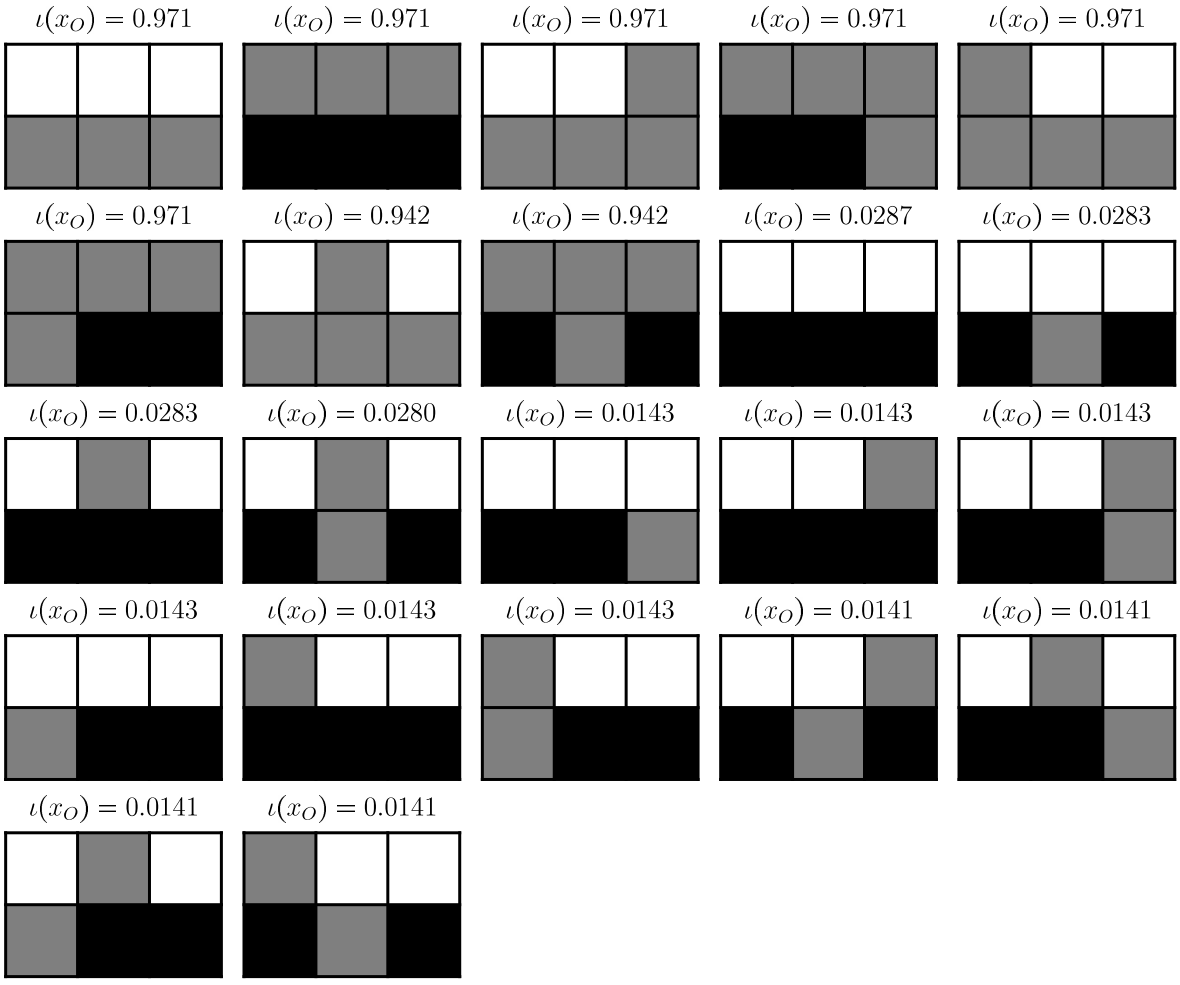


Figure 15. *All distinct completely integrated composite patterns on the first trajectory x_V^1 of MC^ϵ. The value of complete local integration is indicated above each pattern. See Fig. 9 for colouring conventions.*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 March 2017 doi:10.20944/preprints201703.0171.v1

Peer-reviewed version available at Entropy 2017, 19, , 230; doi:10.3390/e19050230

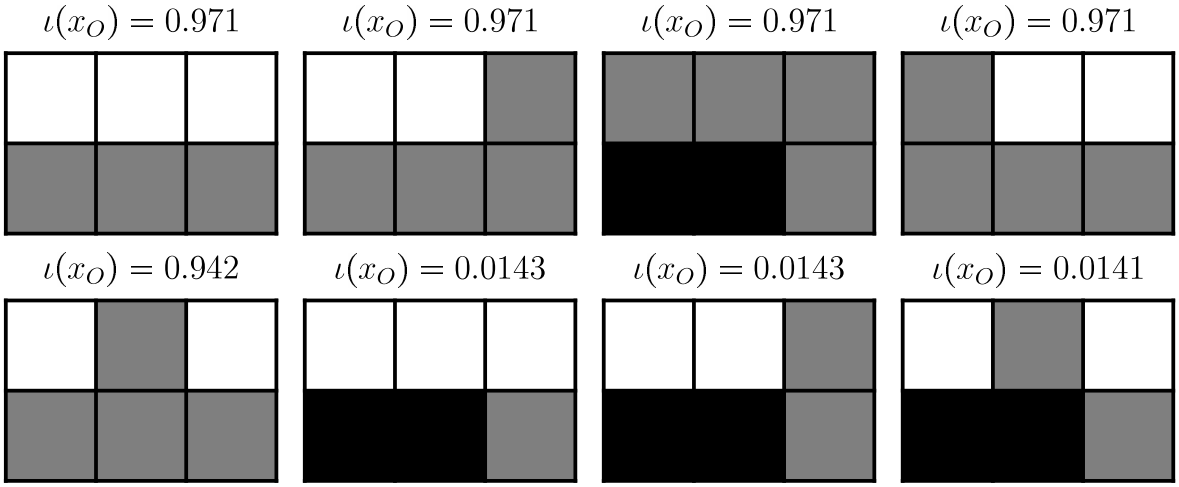


Figure 16. *All distinct completely integrated composite patterns on the second trajectory x_V^2 of MC^ϵ. The value of complete local integration is indicated above each pattern.*

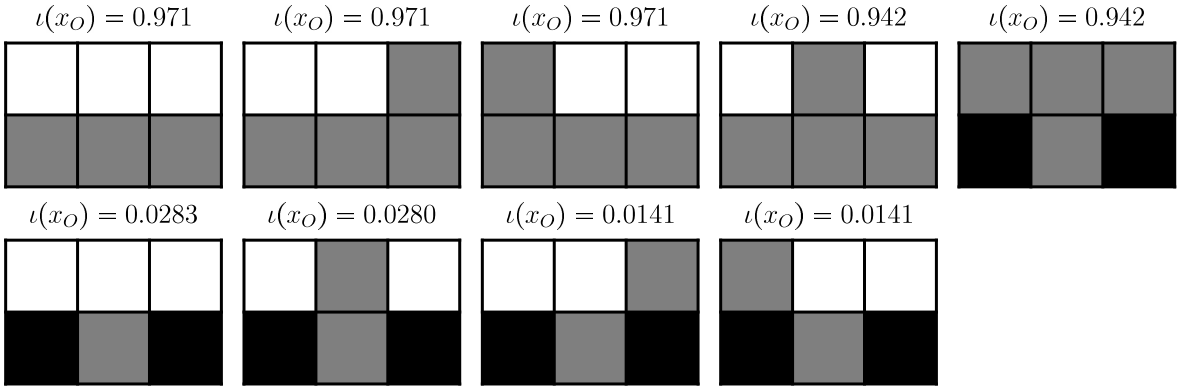


Figure 17. *All distinct completely integrated composite patterns on the third trajectory x_V^3 of MC^ϵ. The value of complete local integration is indicated above each pattern.*

in Figs. 15 to 17. In contrast to the situation of $MC^=$ we now have completely integrated patterns 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 Fig. 10). These are also more than an order of magnitude more integrated than the rest of the completely integrated patterns. This is also true for the other two trajectories.

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 in one-to-one correspondence to the subsets of trajectories that uniquely specify a subset $A \subseteq V$ of a set of random variables $\{X_i\}_{i \in V}$ and leave the complement set $V \setminus A$ of the random variables completely unspecified. This is distinct from arbitrary subsets of trajectories. One of the main target applications of patterns is in time-unrolled Bayesian networks of cellular automata like the one in Fig. 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 [15]. We would also like to investigate whether these spatiotemporal patterns are 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 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.

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. We also calculated the completely locally integrated patterns i.e. the ι-entities of $MC^=$.

These 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 completely integrated patterns can be completely integrated patterns 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 to 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 (Figs. 15 to 17). 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.* [7] 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.

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 [39, 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.

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 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. Currently we are not aware of an analogous 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.

- A first consequence of introducing the logarithm is that we can now formulate the condition of Eq. (18) analogously to an old phrase attributed to Aristotle that "the whole is more than the sum of its parts". In our case this would need to be changed to "the log-probability of the (spatiotemporal) whole is greater than the sum of the log-probabilities of its (spatiotemporal) parts". This can easily be seen by rewriting Eq. (20) as:

$$\mathrm{mi}_\pi(x_O) = \log p_O(x_O) - \sum_{b\in\pi} \log p_b(x_b). \tag{55}$$

- Another side effect of using the logarithm is that we can interpret Eq. (18) in terms of the surprise value (also called information content) $-\log p_O(x_O)$ [40] of the pattern x_O and the surprise value of its parts with respect to any partition π. Rewriting Eq. (20) using properties of the logarithm we get:

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

  Interpreting Eq. (18) 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.
- With respect to hypothesis testing, we can view the product probability $\prod_{b\in\pi} p_b(x_b)$ with respect to partition π as the probability of x_O associated with the hypothesis that the parts x_b are stochastically independent. Let us call this hypothesis $\mathcal{H}_\pi$. Then we can write:

$$p(x_O|\mathcal{H}_\pi) := \prod_{b\in\pi} p_b(x_b). \tag{56}$$

Similarly, we can view the joint probability $p_O(x_O)$ as the probability of x_O under the hypothesis that the full joint probability is needed. Let us write $\mathcal{H}_\mathbf{1}$ for this hypothesis and define accordingly:

$$p(x_O|\mathcal{H}_\mathbf{1}) := p_O(x_O). \tag{57}$$

The occurrence of x_O is then said to provide what is called the "weight of evidence in favour of $\mathcal{H}_\mathbf{1}$" [40] defined by

$$\log \frac{p(x_O|\mathcal{H}_\mathbf{1})}{p(x_O|\mathcal{H}_\pi)} > 0. \tag{58}$$

So in this terminology a completely locally integrated pattern x_O provides evidence in favour of $\mathcal{H}_\mathbf{1}$ compared to *each* hypothesis $\mathcal{H}_\pi$, $\pi \in \mathfrak{L}(O) \setminus \mathbf{1}$ that supposes it is composite of stochastically independent parts.

- In coding theory, the Kraft-McMillan theorem [41] tells us that the optimal length (in a uniquely decodable binary code) of a code word for an event x is $l(x) = -\log p(x)$ if $p(x)$ is the *true* probability of x. If the encoding is not based on the true probability of x but instead on a different probability $q(x)$ then the difference between the optimal code word length and the chosen code word length is

$$-\log q(x) - (-\log p(x)) = \log \frac{p(x)}{q(x)}. \tag{59}$$

Then we can interpret the specific local integration as a difference in code word lengths. Say we want to encode what occurs at the nodes/random variables indexed by O i.e. we encode the random variable X_V. We can encode every event (now a 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 code word lengths between the joint code and the product code with respect to π is then just the specific local integration with respect to π.

Complete local integration then requires that the joint code code word is shorter than all possible product code code words. This means there is no partition with respect to which the product code for the pattern x_O has a shorter code word 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 partitions 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 $\mathcal{I}_\pi(X_O)$ [8,9] with respect to π, i.e.

$$\mathcal{I}_\pi(X_O) := \sum_{x_O\in\mathcal{X}_O} p_O(x_O) \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)} \tag{60}$$

$$= \sum_{x_O\in\mathcal{X}_O} p_O(x_O)\, \mathrm{mi}_\pi(x_O). \tag{61}$$

The multi-information plays a role in measures of complexity and information integration [37]. The generalisation from bipartitions to arbitrary partitions is applied to expectation values similar to the multi-information above in Tononi [36]. 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 the values x_O.

- Using the disintegration theorem (Theorem 5) results in yet another point of view. The theorem states that for each trajectory $x_V \in \mathcal{X}_V$ of a multivariate Markov chain the refinement-free

disintegration hierarchy only contains 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 code word lengths $-\log p_V(x_V)$ for each trajectory x_V. The partitions in the lowest level $\mathfrak{D}_1^{\blacktriangleleft}(x_V)$ in the refinement-free disintegration hierarchy for x_V have minimal specific local integration i.e.

$$\mathrm{mi}_\pi(x_V) = \log \frac{p_V(x_V)}{\prod_{c \in \pi} p_c(x_c)} \tag{62}$$

is minimal among all partitions. At the same time these partitions are the finest partitions that achieve this low specific local integration. This implies on the one hand that the code word lengths of the product codes associated to these partitions are the shortest possible for x_V among all partitions. On the other hand these partitions split up the trajectory in as many parts as possible while generating these shortest code words. In this combined sense the partitions in $\mathfrak{D}_1^{\blacktriangleleft}(x_V)$ generate the "best" product codes for the particular trajectory x_V.

Note that the *expected code word length* of the product code:

$$\sum_{x_V \in \mathcal{X}_V} p_V(x_V)(-\log \prod_{c \in \pi} p_c(x_c)) \tag{63}$$

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 code word 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 [41] 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 code word for x_V to be the shortest possible. Then we would use the product codes of partitions in the refinement-free disintegration hierarchy because those combine exactly these two conditions. The pseudo-stochastically-independent parts would then be the blocks of these partitions which according to the disintegration theorem are exactly the ι-entities occurring in x_V.

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. 42]. 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 code word length). It is possibly more of a naive code that does not require deep understanding of the dynamical system but instead can be learned fast and works. The language of physics for example might be more optimal in the sense of shortest expected code word lengths reflecting a desire to communicate efficiently about all counterfactual possibilities as well.

6. Conclusions

We have presented a formal investigation into a criterion designed to distinguish arbitrary patterns in Bayesian networks from entities. Patterns were also formally defined and their difference to arbitrary subsets of the set of trajectories of a Bayesian network was established.

The entity criterion is based on the measure of specific local integration (SLI) for which we constructively proved the least upper bound and presented a candidate for the greatest lower bound. The measure of the criterion itself is complete local integration. 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 and the candidate for the greatest lower bound of SLI is also a candidate for the greatest lower bound of CLI.

We then 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.

Finally, we presented the disintegration hierarchy and calculated the ι-entities for some simple example systems. We established that the ι-entities:

- correspond to individual 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.

A more intriguing result is that parts of entities can again be entities. Whether this together with the disintegration levels hints at the capacity of the refinement-free disintegration hierarchy to describe entities on different scales remains to be seen however.

We have also noted that the computational burden of calculating the ι-entities seems prohibitive so that formal proofs seem to be the way forward until significant advances in relevant computations or approximations are available.

One strand of research connected to ι-entities is into the definition of perception and action for general spatiotemporal entities. In combination with the ι-entities this can lead to a fully formal definition of agents within (dynamic) Bayesian networks. Another and more ambitious future goal is to investigate whether ι-entities within a dynamical Bayesian network can be shown to exhibit

Darwinian evolution. This would pave the way for a formal theory of the emergence of evolution and possibly the origin of life in dynamical systems.

Acknowledgments: Part of the research was performed during M.B.'s time as a JSPS International Research Fellow and as an ELSI Origins Network (EON) long term visitor. D.P. was supported in part by the H2020-641321 socSMCs FET Proactive project.

Author Contributions: M.B., T.I., and D.P. conceived the problem and the measures, and wrote the paper; M.B. proved the theorems and conceived and calculated the examples.

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 A.1 (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*

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

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

$$\delta_x(\bar{x}) := \begin{cases} 1 & \text{if } x \in \{\bar{x}\}, \\ 0 & \text{else.} \end{cases} \tag{A.2}$$

$$= \begin{cases} 1 & \text{if } x = \bar{x}, \\ 0 & \text{else.} \end{cases} \tag{A.3}$$

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

$$p(y|x) = \delta_{f(x)}(y), \tag{A.4}$$

then

$$p(y) = \sum_x p_Y(y|x) p_X(x) \tag{A.5}$$

$$= \sum_x \delta_{f(x)}(y) p_X(x) \tag{A.6}$$

$$= \sum_x \delta_x(f^{-1}(y)) p_X(x) \tag{A.7}$$

$$= \sum_{x \in f^{-1}(y)} p_X(x) \tag{A.8}$$

$$= p_X(f^{-1}(y)). \tag{A.9}$$

Appendix B. 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:

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

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 Eq. (B.1). 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 Eq. (B.1). 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 B.1. *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:*

$$p_V(x_V) = \prod_{i\in V} p(x_i|x_{\mathrm{pa}(i)}). \tag{B.2}$$

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 Eq. (B.2) becomes Eq. (B.1) which always holds.

Definition B.2 (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.*

Remark:

- 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 B.3 (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)*

$$\sum_{x_i \in \mathcal{X}_i} p_i(x_i|x_{\mathrm{pa}(i)}) = 1. \tag{B.3}$$

Remark:

- We could define the set of all mechanisms to formally also include the mechanisms of the nodes without parents V_0. However in practice it makes sense to separate the nodes without parents as those that we choose an initial probability distribution over (similar to a boundary condition) which is then turned into a probability distribution p_V over the entire Bayesian network $\{X_i\}_{i\in V}$ via Eq. (B.2). Note that in Eq. (B.2) 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 Eq. (B.2):

$$p_V(x_V) = \prod_{i \in V\setminus V_0} p_i(x_i|x_{\mathrm{pa}(i)}) \prod_{j \in V_0} p_j(x_j). \tag{B.4}$$

Appendix B.1. Deterministic Bayesian networks

Definition B.4 (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*

$$p_i(x_i|x_{\mathrm{pa}(i)}) = \delta_{f_i(x_{\mathrm{pa}(i)})}(x_i) = \begin{cases} 1 & \text{if } x_i = f_i(x_{\mathrm{pa}(i)}), \\ 0 & \text{else.} \end{cases} \tag{B.5}$$

Definition B.5 (Deterministic Bayesian network)**.** *A Bayesian network* $\{X_i\}_{i\in V}$ *is deterministic if all its mechanisms are deterministic.*

Theorem B.1. *Given a deterministic Bayesian network* $\{X_i\}_{i\in V}$ *there exists a function* $f_{V\setminus V_0} : \mathcal{X}_{V_0} \to \mathcal{X}_{V\setminus V_0}$ *which given a value* x_{V_0} *of the random variables without parents* X_{V_0} *returns the value* $x_{V\setminus V_0}$ *fixing the values of all remaining random variables in the network.*

Proof. According to Eq. (B.2), the definition of conditional probabilities, and using the deterministic mechanisms we have:

$$p_{V\setminus V_0}(x_{V\setminus V_0}|x_{V_0}) = \prod_{i \in V\setminus V_0} p_i(x_i|x_{\mathrm{pa}(i)}) \tag{B.6}$$

$$= \prod_{i \in V\setminus V_0} \delta_{f_i(x_{\mathrm{pa}(i)})}(x_i). \tag{B.7}$$

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 B.2 (Pattern probability in a deterministic Bayesian network)**.** *Given a deterministic Bayesian network (Definition B.5) 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:*

$$p_A(x_A) = \frac{N(x_A)}{|\mathcal{X}_{V_0}|} \tag{B.8}$$

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 B.1) 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

$$\begin{aligned}
p_A(x_A) &= \sum_{\bar{x}_B} p_V(x_A, \bar{x}_B) && \text{(B.9)}\\
&= \sum_{\bar{x}_{B_0}, \bar{x}_{B_r}} p_V(x_{A_r}, \bar{x}_{B_r} | x_{A_0}, \bar{x}_{B_0}) p_{V_0}(x_{A_0}, \bar{x}_{B_0}) && \text{(B.10)}\\
&= \sum_{\bar{x}_{B_0}, \bar{x}_{B_r}} \delta_{f_{V\setminus V_0}(x_{A_0}, \bar{x}_{B_0})}(x_{A_r}, \bar{x}_{B_r}) p_{V_0}(x_{A_0}, \bar{x}_{B_0}) && \text{(B.11)}\\
&= \sum_{\bar{x}_{B_r}} \sum_{\{\bar{x}_{B_0} : (x_{A_0}, \bar{x}_{B_0}) \in f^{-1}_{V\setminus V_0}(x_{A_r}, \bar{x}_{B_r})\}} p_{V_0}(x_{A_0}, \bar{x}_{B_0}) && \text{(B.12)}\\
&= \frac{1}{|\mathcal{X}_{V_0}|} \sum_{\bar{x}_{B_r}} |\{\bar{x}_{B_0} \in \mathcal{X}_{B_0} : (x_{A_0}, \bar{x}_{B_0}) \in f^{-1}_{V\setminus V_0}(x_{A_r}, \bar{x}_{B_r})\}| && \text{(B.13)}\\
&= \frac{1}{|\mathcal{X}_{V_0}|} N(x_A) && \text{(B.14)}
\end{aligned}$$

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 B.2. Proof of Theorem 2

Proof. Follows by replacing the probabilities $p_O(x_O)$ and $p_b(x_b)$ in Eq. (20) with their deterministic expressions from Theorem B.2, i.e. $p_A(x_A) = N(X_A)/|\mathcal{X}_{V_0}|$. Then:

$$\begin{aligned}
\mathrm{mi}_\pi(x_O) :&= \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)} && \text{(B.15)}\\
&= \log \frac{\frac{N(x_O)}{|\mathcal{X}_{V_0}|}}{\prod_{b\in\pi} \frac{N(x_b)}{|\mathcal{X}_{V_0}|}} && \text{(B.16)}\\
&= \log \frac{\frac{N(x_O)}{|\mathcal{X}_{V_0}|}}{|\mathcal{X}_{V_0}|^{-|\pi|} \prod_{b\in\pi} N(x_b)} && \text{(B.17)}\\
&= \log \frac{|\mathcal{X}_{V_0}|^{|\pi|-1} N(x_O)}{\prod_{b\in\pi} N(x_b)} && \text{(B.18)}\\
&= (|\pi|-1)\log|\mathcal{X}_{V_0}| - \log \frac{\prod_{b\in\pi} N(x_b)}{N(x_O)}. && \text{(B.19)}
\end{aligned}$$

□

Appendix C. Proof of Theorem 1

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

We first show that if there exists $A \subseteq V$ with $\mathcal{D}_A \subset \mathcal{X}_V$ 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 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

$$\tilde{x}_B = (\tilde{x}_{B\setminus A}, \tilde{x}_{B\cap A}). \tag{C.1}$$

If $B \cap A \neq \emptyset$ 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 $\mathcal{T}(\bar{x}_A) \notin \mathcal{T}(\tilde{x}_B)$ so $\mathcal{D} \notin \mathcal{T}(\tilde{x}_B)$. Next choose $\tilde{x}_{B\cap A} \neq x_A$ but then $\mathcal{T}(x_A) \notin \mathcal{T}(\tilde{x}_B)$ so also $\mathcal{D} \notin \mathcal{T}(\tilde{x}_B)$. 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 $\mathcal{D}$. 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, if there exists no $A \subseteq V$ with $\mathcal{D}_A \subset \mathcal{X}_V$ and $|\mathcal{D}_A| > 1$, then for each $C \subseteq V$ either $\mathcal{D}_A = \mathcal{X}_A$ or $|\mathcal{D}_A| = 1$. 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 $A \cup B\emptyset$ by assumption $\mathcal{D}_A = \mathcal{X}_A$ 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 $\mathcal{T}(\tilde{x}_B) = \mathcal{D}$. □

Remark:

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

$$\mathcal{D} := \{x_A \cup \bar{x}_A\} = \{(0,0),(1,1)\}. \tag{C.2}$$

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

$C\subseteq V$	x_C	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)	$\{(0,1),(1,1)\}$
$\{1,2\}$	$(0,0)$	$\{(0,0)\}$
	$(0,1)$	$\{(0,1)\}$
	$(1,0)$	$\{(1,0)\}$
	$(1,1)$	$\{(1,1)\}$

(C.3)

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

Appendix D. Bounds

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 D.1 (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:*

$$\neg(x_O):=\{\bar{x}_O\in\mathcal{X}_O:\forall i\in O,\bar{x}_i\neq x_i\}. \tag{D.1}$$

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 D.1 (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*

$$\mathrm{mi}_\pi(x_O)=-(n-1)\log q. \tag{D.2}$$

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

$$\begin{aligned}
\mathrm{mi}_\pi(x_O) &= \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)} && \text{(D.3)}\\
&= \log \frac{p_O(x_O)}{\prod_{b\in\pi} \sum_{\bar{x}_{O\setminus b}} p_O(x_b, \bar{x}_{O\setminus b})} && \text{(D.4)}\\
&= \log \frac{p_O(x_O)}{\prod_{b\in\pi} \left(p_O(x_O) + \sum_{\bar{x}_{O\setminus b} \neq x_{O\setminus b}} p_O(x_b, \bar{x}_{O\setminus b})\right)} && \text{(D.5)}\\
&= \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_O(x_O)} && \text{(D.6)}\\
&= \log \frac{p_O(x_O)}{p_O(x_O)^n} && \text{(D.7)}\\
&= -(n-1)\log q. && \text{(D.8)}
\end{aligned}$$

□

Remark:

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

Appendix D.1. Proof of Theorem 3

Proof. ad 1 By Definition 7 we have

$$\mathrm{mi}_\pi(x_O) = \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)}. \tag{D.9}$$

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

$$\begin{aligned}
p_b(x_b) &= \sum_{\bar{x}_{O\setminus b}} p_O(x_b, \bar{x}_{O\setminus b}) && \text{(D.10)}\\
&= p_O(x_O) + \sum_{\bar{x}_{O\setminus b} \neq x_{O\setminus b}} p_O(x_b, \bar{x}_{O\setminus b}) && \text{(D.11)}\\
&\geq p_O(x_O). && \text{(D.12)}
\end{aligned}$$

Plugging this into Eq. (D.9) for every $p_b(x_b)$ we get

$$\begin{aligned}
\mathrm{mi}_\pi(x_O) &= \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)} && \text{(D.13)}\\
&\leq \log \frac{p_O(x_O)}{p_O(x_O)^{|\pi|}} && \text{(D.14)}\\
&= -(|\pi| - 1)\log p_O(x_O). && \text{(D.15)}
\end{aligned}$$

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 D.1.

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

$$\log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)} = -(|\pi|-1)\log p_O(x_O) \tag{D.16}$$

$$\Leftrightarrow \quad \log \frac{p_O(x_O)}{\prod_{b\in\pi} p_b(x_b)} = \log \frac{p_O(x_O)}{p_O(x_O)^{|\pi|}} \tag{D.17}$$

$$\Leftrightarrow \quad \prod_{b\in\pi} p_b(x_b) = p_O(x_O)^{|\pi|}, \tag{D.18}$$

and because $p_b(x_b) \geq p_O(x_O)$ (Eq. (D.12)) 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 Eq. (D.11) we then see that

$$p_b(x_b) = p_O(x_O) + \sum_{\bar{x}_{O\setminus b} \neq x_{O\setminus b}} p_O(x_b, \bar{x}_{O\setminus b}) \tag{D.19}$$

$$> p_O(x_O). \tag{D.20}$$

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

□

Appendix D.2. Proof of Theorem 4

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 D.1 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:

$$p_O(\bar{x}_O) := \begin{cases} q & \text{if } \bar{x}_O = x_O, \\ \frac{1-q-d}{\sum_{b\in\pi} |\neg(x_b)|} & \text{if } \exists c \in \pi \text{ s.t. } \bar{x}_{O\setminus c} = x_{O\setminus c} \wedge \bar{x}_c \neq x_c, \\ \frac{d}{|\neg(x_O)|} & \text{if } \bar{x}_O \in \neg(x_O), \\ 0 & \text{else.} \end{cases} \tag{D.21}$$

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 $b \in \pi$

$$\sum_{\bar{x}_{O\setminus b}\neq x_{O\setminus b}} p_O(x_b,\bar{x}_{O\setminus b}) = \sum_{c\in\pi\setminus b}\sum_{\bar{x}_c\neq x_c} p_O(x_b,x_{O\setminus(b\cup c)},\bar{x}_c) \tag{D.22}$$

$$= \sum_{c\in\pi\setminus b}\sum_{\bar{x}_c\neq x_c} \frac{1-q-d}{\sum_{b\in\pi}|\neg(x_b)|} \tag{D.23}$$

$$= \sum_{c\in\pi\setminus b}\sum_{\bar{x}_c\neq x_c} \frac{1-q-d}{m|\pi|} \tag{D.24}$$

$$= \sum_{c\in\pi\setminus b} \frac{1-q-d}{m|\pi|}|\neg(x_c)| \tag{D.25}$$

$$= \frac{|\pi|-1}{|\pi|}(1-q-d) \tag{D.26}$$

Plug this into the SLI definition:

$$\mathrm{mi}_\pi(x_O) = \log\frac{p_O(x_O)}{\prod_{b\in\pi}p_b(x_b)} \tag{D.27}$$

$$= \log\frac{q}{\prod_{b\in\pi}q+\sum_{\bar{x}_{O\setminus b}\neq x_{O\setminus b}}p_O(x_b,\bar{x}_{O\setminus b})} \tag{D.28}$$

$$= \log\frac{q}{\prod_{b\in\pi}q+\frac{|\pi|-1}{|\pi|}(1-q-d)} \tag{D.29}$$

$$= \log\frac{q}{\left(q+\frac{|\pi|-1}{|\pi|}(1-q-d)\right)^{|\pi|}} \tag{D.30}$$

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

$$\mathrm{mi}_\pi(x_O) = \log\frac{q}{\left(1-\frac{1-q}{|\pi|}\right)^{|\pi|}}. \tag{D.31}$$

Then we can use Bernoulli's inequality[10] to prove that this is negative for $0<q<1$ and $|\pi|\geq 2$. Bernoulli's inequality is

$$(1+x)^n \geq 1+nx \tag{D.32}$$

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

$$\left(1-\frac{1-q}{|\pi|}\right)^{|\pi|} > q \tag{D.33}$$

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

Remarks:

- In order to achieve the negative SLI of Eq. (D.31) the requirement is only that Eq. (D.26) 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 such that Eq. (D.26) still holds. This is not necessary for our purpose here as we only want to show the existence of patterns achieving the negative value.

[10] The authors thank von Eitzen [43] for pointing this out. An example reference for Bernoulli's inequality is Bullen [44].

Bibliography

1. Gallois, A. Identity Over Time. In *The Stanford Encyclopedia of Philosophy*, Summer 2012 ed.; Zalta, E.N., Ed.; 2012.
2. Grand, S. *Creation: Life and How to Make It*; Harvard University Press, 2003.
3. Pascal, R.; Pross, A. Stability and its manifestation in the chemical and biological worlds. *Chemical Communications* **2015**, *51*, 16160–16165.
4. Orseau, L.; Ring, M. Space-Time Embedded Intelligence. In *Artificial General Intelligence*; Bach, J.; Goertzel, B.; Iklé, M., Eds.; Number 7716 in Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2012; pp. 209–218.
5. Barandiaran, X.E.; Paolo, E.D.; Rohde, M. Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-temporality in Action. *Adaptive Behavior* **2009**, *17*, 367–386.
6. Legg, S.; Hutter, M. Universal Intelligence: A Definition of Machine Intelligence. *arXiv:0712.3329 [cs]* **2007**. arXiv: 0712.3329.
7. Biehl, M.; Ikegami, T.; Polani, D. Towards information based spatiotemporal patterns as a foundation for agent representation in dynamical systems. Proceedings of the Artificial Life Conference 2016. The MIT Press, 2016, pp. 722–729.
8. McGill, W.J. Multivariate information transmission. *Psychometrika* **1954**, *19*, 97–116.
9. Amari, S.I. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory* **2001**, *47*, 1701–1711.
10. Ay, N. Information geometry on complexity and stochastic interaction. *MPI MiS Preprint* **2001**, *95/2001*. Available at: http://www.mis.mpg.de/publications/preprints/2001/prepr2001-95.html.
11. Lizier, J.T. *The Local Information Dynamics of Distributed Computation in Complex Systems*; Springer Science & Business Media, 2012.
12. Tononi, G.; Sporns, O. Measuring information integration. *BMC Neuroscience* **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. Oizumi, M.; Albantakis, L.; Tononi, G. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Comput Biol* **2014**, *10*, e1003588.
15. Beer, R.D. Characterizing autopoiesis in the game of life. *Artificial Life* **2014**, *21*, 1–19.
16. Krakauer, D.; Bertschinger, N.; Olbrich, E.; Ay, N.; Flack, J.C. The Information Theory of Individuality. *arXiv:1412.2447 [q-bio]* **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. *Physical Review E* **2006**, *73*, 036104.
19. Wolfram, S. Computation theory of cellular automata. *Communications in mathematical physics* **1984**, *96*, 15–57.
20. Grassberger, P. Chaos and diffusion in deterministic cellular automata. *Physica D: Nonlinear Phenomena* **1984**, *10*, 52–58.
21. Hanson, J.E.; Crutchfield, J.P. The attractor—basin portrait of a cellular automaton. *Journal of Statistical Physics* **1992**, *66*, 1415–1462.
22. Pivato, M. Defect particle kinematics in one-dimensional cellular automata. *Theoretical Computer Science* **2007**, *377*, 205–228.
23. Lizier, J.T.; Prokopenko, M.; Zomaya, A.Y. Local information transfer as a spatiotemporal filter for complex systems. *Physical Review 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: An Interdisciplinary Journal of Nonlinear Science* **2011**, *21*, 037104.
25. Friston, K. Life as we know it. *Journal of The Royal Society Interface* **2013**, *10*.
26. Balduzzi, D. Detecting emergent processes in cellular automata with excess information. *Advances in Artificial Life, ECAL* **2011**, *abs/1105.0158*.
27. Hoel, E.P.; Albantakis, L.; Tononi, G. Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences of the United States of America* **2013**, *110*, 19790–19795.

28. Hoel, E.P.; Albantakis, L.; Marshall, W.; Tononi, G. Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neuroscience of Consciousness* **2016**, *2016*.
29. Noonan, H.; Curtis, B. Identity. In *The Stanford Encyclopedia of Philosophy*, Summer 2014 ed.; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University, 2014.
30. Grätzer, G. *Lattice Theory: Foundation*, 2011 edition ed.; Springer: Basel ; New York, 2011.
31. Ceccherini-Silberstein, T.; Coornaert, M. Cellular Automata and Groups. In *Encyclopedia of Complexity and Systems Science*; Ph.D, R.A.M., Ed.; Springer New York, 2009; pp. 778–791. DOI: 10.1007/978-0-387-30440-3_52.
32. Busic, A.; Mairesse, J.; Marcovici, I. Probabilistic cellular automata, invariant measures, and perfect sampling. *arXiv:1010.3133 [cs, math]* **2010**. arXiv: 1010.3133.
33. Beer, R.D. The cognitive domain of a glider in the game of life. *Artificial Life* **2014**, *20*, 183–206.
34. Beer, R.R. Autopoiesis and Enaction in the Game of Life. The MIT Press, 2016, pp. 13–13.
35. Hawley, K. Temporal Parts. In *The Stanford Encyclopedia of Philosophy*, Winter 2015 ed.; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University, 2015.
36. Tononi, G. An information integration theory of consciousness. *BMC Neuroscience* **2004**, *5*, 42.
37. Ay, N. Information Geometry on Complexity and Stochastic Interaction. *Entropy* **2015**, *17*, 2432–2458.
38. Pemmaraju, S.; Skiena, S. *Computational Discrete Mathematics: Combinatorics and Graph Theory with Mathematica* ®; Cambridge University Press, 2009.
39. Bruijn, N.G.d. *Asymptotic Methods in Analysis*, dover ed edition ed.; Dover Publications: New York, 2010.
40. MacKay, D.J. *Information theory, inference and learning algorithms*; Cambridge university press, 2003.
41. Cover, T.M.; Thomas, J.A. *Elements of information theory*; Wiley-Interscience: Hoboken, N.J., 2006.
42. Kolchinsky, A.; Rocha, L.M. Prediction and modularity in dynamical systems. *Advances in Artificial Life, ECAL* **2011**, pp. 423–430.
43. 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, [http://math.stackexchange.com/q/1974262]. URL:http://math.stackexchange.com/q/1974262 (version: 2016-10-18).
44. Bullen, P.S. *Handbook of Means and Their Inequalities*; Springer Science & Business Media, 2003.