Comparison between Different Methods of Level Identification

Oliver Pfante
Nils Bertschinger
Eckehard Olbrich
Nihat Ay
Jürgen Jost

SFI WORKING PAPER: 2013-11-035

SANTA FE INSTITUTE

COMPARISON BETWEEN DIFFERENT METHODS OF LEVEL IDENTIFICATION

OLIVER PFANTE[1,A)], NILS BERTSCHINGER[1,B)], ECKEHARD OLBRICH[1,C)], NIHAT AY[1,2,D)], JÜRGEN JOST[1,2,E)]

Abstract. Levels of a complex system are characterized by the fact that they admit a closed functional description in terms of concepts and quantities intrinsic to that level. Several ideas have been come up so far in order to make the notion of a closed description precise. In this paper we present four of these approaches and investigate their mutual relationships. Our study is restricted to the case of discrete dynamical systems where the different levels are linked by a coarse-graining of variables and states of the system.

Keywords. Lumpability; aggregated Markov chains; aggregation of variables; aggregated linear dynamics; state space reduction; coarsegraining.

1. Introduction

Classical scientific models often derive their success from providing a closed description of some section of reality. The Schrödinger equation, for instance, describes the quantum mechanical behavior of the hydrogen atom without being concerned with the fact that the proton constituting its nucleus is not elementary, but is composed of quarks. However, the mechanism of quark confinement makes a consistent description at the atomic level possible. Likewise, Newtonian dynamics models the sun, the planets and their moons as point masses and can then capture their dynamics without having to consider their rather complicated internal structure in terms of solid state physics or fluid or gas dynamics. In biology, the Hodgkin-Huxley equations and their variants operate successfully at the cellular level without needing the molecular details of the ion channels. The Wright-Fisher and other models of population genetics express genetic drift in large populations in terms of Fokker-Planck type equations without the complicated details of the mechanisms of genetic inheritance at the DNA level.

However, such a reducing scheme is not always easily available. Neuroscience or cell biology provide examples where it may not even be clear what the appropriate levels are. And even in physics, the transition from the atomic or molecular scale to a hydrodynamic description in the case of non-Newtonian fluids or phenomenological descriptions in materials science provide some of the most difficult problems that science faces.

The field of complex systems is concerned, in particular, with systems comprising several different levels. Their hierarchy might be relatively clear, as in biology where we have macromolecules, cells, tissues, organs, organisms, populations, and ecosystems, or it may be rather opaque as it seems to be the case in the cognitive neurosciences. In particular, there might be processes that are best understood on intermediate levels that are not a priori known. This raises the question how we can identify such a level given some microscopic description of a system, or better, what are the criteria that a level of a complex system should satisfy. In view of the above, one should ask for some kind of closed description of the dynamics taking place at that level, but admit at the same time that in systems with interacting levels, this cannot be perfectly realized. Nevertheless, when we have formalized such an idealization, this should also provide us with tools to quantify the deviations from the idealized scheme.

Closedness here would mean that the level admits a description that is self-consistent in the sense that it does not need to perpetually draw information from or about the lower-level states. This, however, can be made precise in different ways, and there exist at least two different proposals for what should be considered as closed in the literature:

1. One notion was called computational reducibility in [9]. It means that in order to predict the behavior on a coarse grained level one can use the dynamics on the coarse grained level and does not have to simulate the full microscopic dynamics. Formally this corresponds to a commutativity condition between the coarse graining and the dynamics.
2. The other notion was proposed by Shalizi and Moore [12] where they identify aggregated, averaged or coarse-grained states as "macrostate" of the system if their dynamics is Markovian in this state space. This is usually only approximately fulfilled. In this case one needs to measure the deviation from Markovianity.

Efforts [7] were undertaken by Görnerup and Jacobi to give a measure from the deviation of the coarse-grained macro process to be Markovian.
3. Görnerup and Jacobi investigate in [8] and [6] possible methods for finding lumpings - i. e. the states are combined with respect to a partition $A = \{A_1, \dots, A_n\}$ - such that the macro process is Markovian again no matter what choice is made for the starting vector. If this independence holds one calls the process *strongly lumpable* with respect to this partition. However, there are some interesting theoretical considerations, elaborated in the book of Kemeny and Snell [10] whose outline is based on [4], when one requires only that at least one starting vector leads to a Markov chain. When this is the case the process is called *weakly lumpable with respect to the partition A*.

On the basis of these previous results, we started thinking about possible level identification methods, and the corresponding closure measures and how they are related to each other. For the beginning, we have restricted ourselves to a discrete setting where the state spaces are finite, time is discrete, and all coarse-grainings or dynamics are assumed to be linear. The reasons choosing the finite setup are technical ones: the aggregation method described by Görnerup and Jacobi in [8] and [6] is only applicable in the finite setting. Generalizing it to state spaces with infinitely many elements is the issue of a proceeding paper [1]. Furthermore, the information quantities used in this paper, like entropy and mutual information, may diverge or even ill defined for random variables on infinite state spaces. Also these obstacles are circumvented by means of a finite setup.
Within the described discrete setting we have uncovered two further conditions one might want to be fulfilled if one is talking about a closed level. First, taking up an issue already alluded to above, we look at the information flow which might appear if one switches from the lower, microscopic level description to the upper one. In this context a coarse-grained level is informationally closed if there is no informational flow from the lower to the upper level in the sense that knowing the state in the lower level description does not allow for a better prediction of the state on the upper, coarse grained, level. Secondly, we introduced a commutativity condition which essentially states that it does not matter whether we perform the aggregation first, and then observe the upper process, or we observe the process on the microstate level, and then lumping together the states.
This then calls for an investigation of the relationships between those different ways of level descriptions, i.e., Markovianity of the upper level, vanishing information flow and commutativity. In particular, we shall see that the characterization proposed by Görnerup and Jacobi [8] is different from the three other descriptions and establishes a fourth level identification method in its own right.

The setting investigated here is very general. In particular, we make no specific assumptions about the transition from the lower to the higher level (apart from the technical assumption of linearity which we hope to abandon in subsequent work). That transition could, for instance, either occur by averaging, that is where fluctuations at the micro level might cancel at the macro level, or by aggregation where the elements at the micro level behave in some coordinated fashion that builds up collective effects at the macro level. The comparison between deterministic and stochastic dynamics will provide crucial insights.

In technical terms, we assume that X, X' form a Markov process, with transition kernel ϕ, which can be observed at the higher level $\widehat{X}, \widehat{X}'$ in a lossy fashion. The higher level could result from averaging or aggregating the lower level. We assume that all variables live in discrete state spaces. Thus, usually we think of $\widehat{X}$ as being a coarse-graining of X given by an observation map π.

$$\begin{array}{ccc} \widehat{X} & \overset{\psi}{\dashrightarrow} & \widehat{X}' \\ {\scriptstyle\pi}\uparrow & & \uparrow{\scriptstyle\pi} \\ X & \xrightarrow{\phi} & X' \end{array}$$

FIGURE 1. Basic setup of a low level process that can be observed via coarse-graining.

Due to this diagram the already listed ideas of level identification can be summarized as follows.

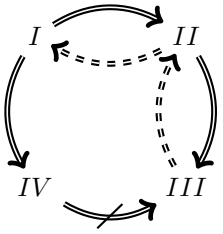
I **Informational closure**: The higher process is informationally closed, i.e. there is no information flow from the lower to the higher level. Knowledge of the microstate will not improve predictions of the macrostate.

II **Observational commutativity**: It makes no difference whether we perform the aggregation first, and then observe the upper process, or we observe the process on the microstate level, and then lumping together the states.

III **Commutativity**: There exists a transition kernel ψ such that the diagram Fig. (1) commutes.

IV **Markovianity**: $\widehat{X}, \widehat{X}'$ forms again a Markov process.

The third idea of a level identification was studied by Görnerup and Jacobi in [8], even though they indicated to study the forth one. The third idea indeed implies the forth one in certain cases, for instance if the coarse-graining is deterministic, but is not equivalent to it, as an counterexample given in this paper shows. In addition to this difference we prove the following implications and inter dependencies between the four ideas.



where the arrows in the diagram encode three different kind of inferences:

⟹ represents an implication which holds true in general.

⇢ represents an implication which holds true if the coarse-graining is deterministic.

⇏ indicates that there is no implication at all, even in the case when the coarse-graining is deterministic.

Furthermore, we could prove that observational commutativity forces the aggregation to be deterministic if the upper process is reversible, and in the case of a soft (i.e. probabilistic) aggregation the third idea of commutativity really differs from the one of observational commutativity.

The outline of the paper is as follows. In sec. 2 we introduce and fix notation of the paper. This is necessary because some proofs and lemmas of this paper are rather technical and there is a need of a precise notation. The very short sec. 3 is dedicated to the first aspect of the listed ideas: Informational flow. We introduce the definition there. In sec. 4 we introduce the idea of observational commutativity. Furthermore, we prove that vanishing informational flow always implies observational commutativity of the lower and the upper processes, and that even the opposite holds true if the coarse-graining is deterministic. Additionally, we prove that a macroscopic dynamics with invertible transition kernel resulting from a soft-aggregation can never observationally commute with the original microscopic dynamics. In sec. 5 we represent the approach towards level identification of Görnerup and Jacobi [8] and link it with the other notions in a twofold way. We disprove that in general the existence of a transition kernel ψ such that the diagram Fig. (1) commutes, also implies observational commutativity. We prove that the existence of such a kernel implies also observational commutativity if the coarse-graining is deterministic. Finally, in sec. 6 we reveal the Markovianity condition to be the very weakest condition for level identification because it is implied by a vanishing informational flow but the opposite does not hold true at all, even if the coarse-graining is deterministic. The Bernoulli-shift provides us with an counterexample. A final discussion in sec. 7 evaluates the results of this paper.

2. Notation

Let the random variables X and X' of the lower level in Fig. (1) be defined on a state space $\{x_1, \dots, x_n\}$ of n elements, and the variables $\widehat{X}$ and $\widehat{X}'$ on a state space $\{\widehat{x}_1, \dots, \widehat{x}_m\}$ of cardinality $m \leq n$. The distributions induced by the random variables can be identified with vectors in $\mathbb{R}^n$, or in $\mathbb{R}^m$ respectively. Due to this identification the transition kernel ϕ of Fig. (1) can be represented by a matrix $P = (p_{ij}) \in M(n \times n, \mathbb{R})$, defined by

$$p_{ij} = p(x'_i|x_j) = p\left(X' = x_i | X = x_j\right) , \tag{2.1}$$

for $i, j = 1, \ldots, n$. Similarly, a matrix representation $\Pi = (\pi_{ij}) \in M(m \times n, \mathbb{R})$ of the observation map π is obtained by

$$\pi_{ij} = \pi(\widehat{x}_i|x_j) = \pi\left(\widehat{X} = \widehat{x}_i | X = x_j\right), \tag{2.2}$$

for $i = 1, \ldots, m$ and $j = 1, \ldots n$. $\pi(\widehat{x}_i|x_j)$ express the probability observing the macrostate $\widehat{x}_i$ conditional on the unobserved event $X = x_j$ and is called the likelihood of $X = x_j$ on $\widehat{X} = \widehat{x}_i$, see [11]. If we fix a prior distribution $p = (p_1 = p(x_1), \ldots, p_n = p(x_n))$ of X, we can compute the posterior distribution $\pi^\dagger(x_j|\widehat{x}_i)$ of X as

$$\pi^\dagger_{ji} = \pi^\dagger(x_j|\widehat{x}_i) = \pi^\dagger\left(X = x_j | \widehat{X} = \widehat{x}_i\right) = \begin{cases} \frac{\pi_{ij} p_j}{\widehat{p}_i} & \text{if } \widehat{p}_i \neq 0 \\ 0 & \text{otherwise} \end{cases}, \tag{2.3}$$

for $j = 1, \ldots, n$, and $i = 1, \ldots, m$, with

$$\widehat{p}_i = \sum_{r=1}^{n} \pi_{ir} p_r \,.$$

If $\widehat{p}_i = 0$ we have also $\pi_{ij} p_j = 0$ for all $j = 1, \ldots, n$. Hence the setting $\pi^\dagger_{ji} = 0$ in this case is consistent. We denote with $\Pi^\dagger = (\pi^\dagger_{ji}) \in M(n \times m, \mathbb{R})$ the matrix representation of the posterior distribution Eq. (2.3). Furthermore, the induced transition kernel ψ on the upper level of Fig. (1) can be represented by a matrix $\widehat{P} = (\widehat{p}_{ij}) \in M(m \times m, \mathbb{R})$:

$$\begin{aligned} \widehat{p}_{ij} = p(\widehat{x}'_i|\widehat{x}_j) &= p\left(\widehat{X}' = \widehat{x}_i | \widehat{X} = \widehat{x}_j\right) \\ &= \sum_{r,s=1}^{n} \pi(\widehat{x}'_i|x'_s) p(x'_s|x_r) \pi^\dagger(x_r|\widehat{x}_j) \\ &= \frac{\sum_{r,s=1}^{n} \pi(\widehat{x}'_i|x'_s) p(x'_s|x_r) \pi(\widehat{x}_j|x_r) p(x_r)}{\sum_{r=1}^{n} \pi(\widehat{x}_j|x_r) p(x_r)} \\ &= \frac{\sum_{r,s=1}^{n} \pi_{is} p_{sr} \pi_{jr} p_r}{\sum_{r=1}^{n} \pi_{jr} p_r} \end{aligned} \tag{2.4}$$

where Eq. (2.3) was used in the second step. If we define the matrices

$$\begin{aligned} D &= \operatorname{diag}(\widehat{p}_1, \ldots, \widehat{p}_m) \\ \Lambda &= \operatorname{diag}(p_1, \ldots, p_n) \,. \end{aligned}$$

Then $\Pi^\dagger$ can be rewritten as

$$\Pi^\dagger = \Lambda \Pi^T D^{-1} \,,$$

with $D^{-1} = \operatorname{diag}(\widehat{p}_1^{-1}, \ldots, \widehat{p}_m^{-1})$ where we set $\widehat{p}_l^{-1} = 0$ if $\widehat{p}_l = 0$. Due to this definitions, the transition kernel $\widehat{P}$ in Eq. (2.4) of the upper process ψ can be written as

$$\widehat{P} = \Pi P \Pi^\dagger \,.$$

3. Informational closure

The first idea can be expressed in terms of entropy and mutual information. The entropy of a random variable X with probability mass function $p(x)$ is defined by

$$H(X) = -\sum_x p(x) \log(p(x)) \,.$$

We use logarithms to base 2. The entropy will then be measured in bits. The entropy is a measure of the average uncertainty in the random variable. It is the number of bits on average required to describe the random variable.

Entropy is the uncertainty of a single random variable. We can define conditional entropy $H(X|Y)$ for two random variables X and Y, with probability mass functions $p(x)$ and $p(y)$ and conditional distribution $p(x|y)$, as

$$H(X|Y) = -\sum_y p(y) \sum_x p(x|y) \log(p(x|y)) \,,$$

which is the entropy of a random variable conditional on the knowledge of another random variable. The reduction in uncertainty due to another random variable is called the mutual information

$$I(X:Y) = H(X) - H(X|Y)\,.$$

The mutual information $I(X:Y)$ is a measure of the dependence between the two random variables. It is symmetric in X and Y and always non negative and is equal to zero if and only if X and Y are independent, see [5].

Definition 3.1. We call the dynamics Fig. (1) of the upper level description informationally closed if

$$I(\widehat{X}' : X|\widehat{X}) = 0\,. \tag{3.1}$$

with the conditional mutual information

$$I(\widehat{X}' : X|\widehat{X}) = H(\widehat{X}'|\widehat{X}) - H(\widehat{X}'|X,\widehat{X}) = H(\widehat{X}'|\widehat{X}) - H(\widehat{X}'|X)$$

where the last equality follows from the fact that $\widehat{X}$ is fully determined by X.

Due to this definition a level description is closed if the variables $\widehat{X}'$ and X are independent given the variable $\widehat{X}$.
This is in the spirit of informational closure that we discussed in the context of system-environment distinction [2] and system autonomy [3]. In this case, we would identify $\widehat{X}$ with the system state S and X denotes its environment E. This interpretation is justified if one assumes that the actual state X consists of the system S and its environment E, i.e. $X = (S, E)$. Then we can write:

$$\begin{aligned} I(\widehat{X}' : X|\widehat{X}) &= I(S' : (S,E)|S) \\ &= I(S' : E|S) + \underbrace{I(S' : S|S,E)}_{=0} \end{aligned}$$

4. Observational Commutativity

Another condition we can look at is that the diagram shown in Fig. (1) observationally commutes, i.e. that the induced transition kernel ψ correctly describes the time evolution of the observed process $\widehat{X} \to \widehat{X}'$. In other words, if one compares the actual distributions

$$p(\widehat{x}'_i|x_j) = \sum_{l=1}^{n} \pi(\widehat{x}'_i|x'_l)p(x'_l|x_j) = \sum_{l=1}^{n} \pi_{il}p_{lj}$$

corresponding to the lower path $X \to X' \to \widehat{X}'$ to the hypothetical distributions

$$p_{upper}(\widehat{x}'_i|x_j) = \sum_{r=1}^{m} p(\widehat{x}'_i|\widehat{x}_r)\pi(\widehat{x}_r|x_j) = \sum_{r=1}^{m} \widehat{p}_{ir}\pi_{rj}\,,$$

corresponding to the upper path $X \to \widehat{X} \to \widehat{X}'$, with $p(\widehat{x}'_i|\widehat{x}_r)$ defined as in Eq. (2.4), the following holds.

Definition 4.1. *Observational Commutativity* holds true if there exists an initial distribution $p = (p(x_1), \ldots, p(x_n))$ different from zero such that

$$\sum_{i=1}^{n} p(x_i) D_{KL}\left(p(\widehat{x}'_r|x_i)||p_{upper}(\widehat{x}'_r|x_i)\right) = 0 \tag{4.1}$$

where $D_{KL}(\cdot||\cdot)$ denotes the Kullback-Leibler distance.

The Kullback-Leibler distance or relative entropy for two probability mass functions $p(x)$ and $q(x)$ is defined as

$$D_{KL}(p||q) = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right) = E_p\left(\log\left(\frac{p(X)}{q(X)}\right)\right)\,,$$

the expected logarithm of the likelihood ratio. In the above definition, we use the convention that $0\log(0/0) = 0$ and the convention (based on continuity arguments) that $0\log(0/q) = 0$ and $p\log(p/0) = \infty$. Thus, if there is any symbol x such that $p(x) > 0$ and $q(x) = 0$, then $D_{KL} = \infty$.
One can prove that the Kullback-Leibler distance is always non negative and is zero if and only if $p = q$, see [5]. However, it is not a true distance between distributions since it is not symmetric and does not satisfy the triangle inequality. Nonetheless, it is often useful to think of relative entropy as a "distance" between distributions.

Theorem 4.1. *Eq. (3.1) implies Eq. (4.1). I. e. if the informational flow vanishes, the processes observationally commute.*

Proof. The Kullback-Leibler distance in Eq. (4.1) can be bounded as follows:

$$
\begin{aligned}
&\sum_{i=1}^{n} p(x_i) D_{KL}(p(\widehat{x}'_r|x_i) \| p_{upper}(\widehat{x}'_r|x_i)) \\
=&\sum_{i=1}^{n} p(x_i) \sum_{r=1}^{m} p(\widehat{x}'_r|x_i) \log \frac{p(\widehat{x}'_r|x_i)}{p_{upper}(\widehat{x}'_r|x_i)} \\
=&\sum_{i=1}^{n} p(x_i) \sum_{r=1}^{m} p(\widehat{x}'_r|x_i) \log \frac{p(\widehat{x}'_r|x_i)}{\sum_{s=1}^{m} p(\widehat{x}'_r|\widehat{x}_s)\pi(\widehat{x}_s|x_i)} \\
=&\sum_{i=1}^{n} p(x_i) \sum_{r=1}^{m} p(\widehat{x}'_r|x_i) \left[\log p(\widehat{x}'_r|x_i) - \log \sum_{s=1}^{m} p(\widehat{x}'_r|\widehat{x}_s)\pi(\widehat{x}_s|x_i) \right] \\
\leq&\sum_{i=1}^{n} p(x_i) \sum_{r=1}^{m} p(\widehat{x}'_r|x_i) \left[\log p(\widehat{x}'_r|x_i) - \sum_{s=1}^{m} \pi(\widehat{x}_s|x_i) \log p(\widehat{x}'_r|\widehat{x}_s) \right] \qquad (4.2) \\
=& \sum_{i=1}^{n} p(x_i) \sum_{r=1}^{m} p(\widehat{x}'_r|x_i) \log p(\widehat{x}'_r|x_i) \\
&- \sum_{i=1}^{n} p(x_i) \sum_{r=1}^{m} p(\widehat{x}'_r|x_i) \sum_{s=1}^{m} \pi(\widehat{x}_s|x_i) \log p(\widehat{x}'_r|\widehat{x}_s) \\
=&- H(\widehat{X}'|X) - \sum_{r,s=1}^{m} p(\widehat{x}_s, \widehat{x}'_r) \log p(\widehat{x}'_r|\widehat{x}_s) \\
=&H(\widehat{X}'|\widehat{X}) - H(\widehat{X}'|X) \\
=&I(X : \widehat{X}'|\widehat{X})
\end{aligned}
$$

where we have used Jensen's inequality and the fact that

$$p(\widehat{x}_s, \widehat{x}'_r) = \sum_{i=1}^{m} p(x_i) p(\widehat{x}'_r|x_i) \pi(\widehat{x}_s|x_i) \,.$$

□

Corollary 4.2. *If the coarse-graining of X is deterministic then from Eq. (4.1) follows Eq. (3.1)*

Proof. If the coarse-graining of X is deterministic, there is for every $i = 1, \ldots, n$ exactly one $s_i \in \{1, \ldots, m\}$ such that

$$\pi(\widehat{x}_s|x_i) = \begin{cases} 1 & \text{if } s = s_i \\ 0 & \texttt{else} \end{cases} \,.$$

Hence

$$\log \sum_{s=1}^{m} \pi(\widehat{x}_s|x_i) p(\widehat{x}'_r|\widehat{x}_s) = \log p(\widehat{x}'_r|\widehat{x}_{s_i}) = \sum_{s=1}^{m} \pi(\widehat{x}_s|x_i) \log p(\widehat{x}'_r|\widehat{x}_s) \,,$$

and the inequality Eq. (4.2) turns into an identity. □

In the sequel we want to investigate under what circumstances the upper and lower processes can observationally commute at all. It turns out that observational commutativity, and due to theorem 4.1 also a vanishing information flow, is a rather strict condition. It forces, under a slight additional assumption concerning the matrix representation $\widehat{P}$ (see Eq. (2.4)) of the transition kernel ψ of the upper process, the coarse-graining to be deterministic.
For the rest of this section we restrict the initial distribution vector

$$p = (p_1 = p(x_1), \ldots, p_n = p(x_n))$$

to be an eigenvector of P, the matrix representation of the transition kernel of the lower process (see Eq. (2.1)), with eigenvalue 1. This assumption is reasonable since we want to obtain a stationary process on the upper level. But the matrix representation $\widehat{P}$ of the upper transition kernel is given by $\widehat{P} = \Pi P \Pi^{\dagger}$ where Π is the matrix representation of the observation map π and $\Pi^{\dagger}$ the one of the Bayesian inverse

(see Eq. (2.3)). The entries of $\Pi^\dagger$ depend on the initial distribution p of the lower level, which is Pp in the next step. Hence the assumption $Pp = p$ is reasonable.
In order to prove the mentioned result we need a rather technical lemma, which translates the definition of 4.1 into the language of linear algebra.

Lemma 4.3. *Let $p = (p_1, \dots, p_r, 0, \dots, 0)$ be the initial distribution vector, which is stationary, i. e. $Pp = p$. Then the matrix P has the form*

$$P = \begin{pmatrix} \tilde{P} & V \\ 0 & W \end{pmatrix},$$

with an $r \times r$ matrix $\tilde{P}$, an $r \times n-r$ matrix V, and an $n-r \times n-r$ matrix W. If we define the $m \times r$ matrix $\tilde{\Pi}$, and the $m \times n-r$ matrix U such that

$$\Pi = \begin{pmatrix} \tilde{\Pi} & U \end{pmatrix}.$$

Then $\Pi^\dagger$ reads as

$$\Pi^\dagger = \begin{pmatrix} \tilde{\Pi}^\dagger \\ 0 \end{pmatrix}.$$

Furthermore, we have $\widehat{P} = \tilde{\Pi}\tilde{P}\tilde{\Pi}^\dagger$, and Eq. (4.1) is equivalent to the equation $\widehat{P}\tilde{\Pi} = \tilde{\Pi}\tilde{P}$.

Proof. Let $P = (p_{ij})_{i,j=1,\dots,n}$. Then $p_{ij} \geq 0$ for all $i, j = 1, \dots, n$. From this and $Pp = p$ one obtains that $p_{ij} = 0$ for all $i = r+1, \dots, n$ and $j = 1, \dots, r$. In addition,

$$\pi^\dagger_{ij} = \begin{cases} \dfrac{\pi_{ji} p_i}{\sum_{s=1}^r \pi_{is} p_s} & \text{for } i = 1, \dots, r, \text{ and } j = 1, \dots, m \\ 0 & \text{otherwise} \end{cases}$$

Due to these observations, it is clear that P, Π and $\Pi^\dagger$ have the form as proposed, and

$$\widehat{P} = \Pi P \Pi^\dagger = \tilde{\Pi}\tilde{P}\tilde{\Pi}^\dagger.$$

From Eq. (4.1) we get for all $j = 1, \dots, m$

$$\begin{array}{lll} & \sum_{i=1}^n p_i D_{KL}\left(p(\widehat{x}_j|x_i)||p_{upper}(\widehat{x}_j|x_i)\right) = 0 & \\ \Leftrightarrow & \sum_{i=1}^r p_i D_{KL}\left(p(\widehat{x}_j|x_i)||p_{upper}(\widehat{x}_j|x_i)\right) = 0 & \\ \Leftrightarrow & D_{KL}\left(p(\widehat{x}_j|x_i)||p_{upper}(\widehat{x}_j|x_i)\right) = 0 & \text{for } i = 1, \dots, r \\ \Leftrightarrow & p(\widehat{x}_j|x_i) = p_{upper}(\widehat{x}_j|x_i) & \text{for } i = 1, \dots, r \\ \Leftrightarrow & e_j^T \Pi P e_i = e_j^T \widehat{P} \Pi e_i & \text{for } i = 1, \dots, r \end{array}$$

where the second and the third equivalence follows by the fact that $D_{KL}(\cdot||\cdot) \geq 0$ and equals 0 iff both entries are equal. But the last equations read, in matrix notation, exactly as $\widehat{P}\tilde{\Pi} = \tilde{\Pi}\tilde{P}$. □

The previous lemma shows that we can simplify our investigations in a twofold manner. Firstly, we can assume that all entries of the initial distribution vector are unequal zero. Secondly, the lemma suggests that we should have a closer look onto the commutation relation $\widehat{P}\Pi = \Pi P$ and its implications. The latter needs the following definition.

Definition 4.2. We call a matrix Π *deterministic* iff in every single column there is exactly one entry different from 0 and equal 1. We call the upper process *reversible* iff the matrix representation $\widehat{P}$ of transition kernel ψ has full rank, i. e. iff the matrix $\widehat{P}$ is invertible.

Theorem 4.4. *Let $\widehat{P}\Pi = \Pi P$, the upper process reversible, and all entries p_i of the initial distribution vector p be different from 0. Then Π must be deterministic.*

Proof. From the reversibility of the upper process we get

$$\operatorname{rank} \Pi \geq \operatorname{rank} \Pi P \Pi^\dagger = \operatorname{rank} \widehat{P} = m,$$

i. e. the matrix Π has full rank. Hence for every $s = 1, \dots, m$ there is at least one $i_0 \in \{1, \dots, n\}$ such that $\pi_{si_0} > 0$. This yields

$$\widehat{p}_s = \sum_{i=1}^n \pi_{si} p_i \geq \pi_{si_0} p_{i_0} > 0,$$

for all $s = 1, \ldots, m$. We have

$$\begin{aligned}\Pi P = \widehat{P}\Pi = \Pi P \Pi^\dagger \Pi = \Pi P \Lambda \Pi^T D^{-1} \Pi \\ = \Pi P \sqrt{\Lambda}\sqrt{\Lambda}\Pi^T D^{-1}\Pi\sqrt{\Lambda}\sqrt{\Lambda}^{-1},\end{aligned}$$

with $\sqrt{\Lambda} = \mathrm{diag}(\sqrt{p_1}, \ldots, \sqrt{p_n})$. We define $A = \Pi P\sqrt{\Lambda}$, and $B = \sqrt{\Lambda}\Pi^T D^{-1}\Pi\sqrt{\Lambda}$. Then the above equation reads as $A = AB$, with a symmetric matrix $B = B^T$. Let denote with $Q : \mathbb{R}^n \to \mathbb{R}^n$ the projection onto the domain $\mathrm{dom}\, A = (\ker A)^\perp$ of the endomorphism on $\mathbb{R}^n$ given by the matrix A, with respect to the standard basis of $\mathbb{R}^n$. With Q we also denote the $n \times n$ matrix-representation of the linear map Q with respect to the standard basis of $\mathbb{R}^n$. From the equation $A = AB$ we obtain $A(1 - B) = 0$, hence $\mathrm{im}\,(1 - B) \subset \ker A$. This yields

$$\begin{aligned}Q(1 - B) = 0 \\ \Leftrightarrow QB = Q\,.\end{aligned}$$

Let $q_1, \ldots, q_n$ be an orthogonal basis of $\mathbb{R}^n$ such that

$$\begin{aligned}\mathrm{im}\; Q = \langle q_1, \ldots, q_r\rangle \\ \ker Q = \langle q_{r+1}, \ldots, q_n\rangle\end{aligned}$$

where $\langle\cdot\rangle$ denotes the linear span, and $r = \mathrm{rank}\; Q$. We compute r as

$$r = \dim\,(\mathrm{dom}\; A) = \mathrm{rank}\; A = \mathrm{rank}(\Pi P\sqrt{\Lambda}) = \mathrm{rank}(\Pi P) = \mathrm{rank}\; \widehat{P}\Pi = \mathrm{rank}\; \Pi = m$$

where the fourth equality follows from the bijectivity of $\sqrt{\Lambda}$, and the sixth one by the assumed reversibility of the upper process. For $i \leq m$ and $j \leq n$ we obtain

$$\langle q_i|Bq_j\rangle = \langle Qq_i|Bq_j\rangle = \langle q_i|QBq_j\rangle = \langle q_i|Qq_j\rangle = \delta_{ij} = \langle q_j|Bq_i\rangle \tag{4.3}$$

where the last equality follows from the symmetry $B = B^T$ of B. If we define the orthogonal $n \times n$ matrix V as

$$V = (q_1, \ldots, q_n)\,,$$

the equations (4.3) can be written as

$$V^T B V = \begin{pmatrix} I_{m\times m} & 0 \\ 0 & U \end{pmatrix}$$

where $I_{m\times m}$ denotes the $m \times m$ identity matrix, and U an $(m - n) \times (m - n)$ matrix. Since

$$m = \mathrm{rank}\; \Pi \geq \mathrm{rank}\; B \geq \mathrm{rank}\; QB = \mathrm{rank}\; Q = m\,,$$

we have rank $B = m$ which proves $U = 0$ and therefore $B = Q$. In particular B is also a projection. This yields

$$\begin{aligned}&\sqrt{\Lambda}\Pi^T D^{-1}\Pi\sqrt{\Lambda} = B = B^2 = \sqrt{\Lambda}\Pi^T D^{-1}\Pi\sqrt{\Lambda}\sqrt{\Lambda}\Pi^T D^{-1}\Pi\sqrt{\Lambda} \\ &\Leftrightarrow \Pi^T D^{-1}\Pi = \Pi^T D^{-1}\Pi\Lambda\Pi^T D^{-1}\Pi \\ &\Leftrightarrow D^{-1} = D^{-1}\Pi\Lambda\Pi^T D^{-1} \\ &\Leftrightarrow 1_{m\times m} = \Pi\Pi^\dagger\end{aligned}$$

where the second equivalence follows from the fact that Π has full rank, which implies that the linear map, induced by Π, is surjective, and the one which is induced by Π^T is injective. This identity yields

$$\delta_{rs} = \sum_{i=1}^{n} \frac{\pi_{ri}\pi_{si}p_i}{d_s}\,, \tag{4.4}$$

for all $r, s = 1, \ldots, m$. Since all $\pi_{ij} \geq 0$, for all $i \leq m$ and $j \leq n$, this forces $\pi_{ri}\pi_{si} = 0$ for every single $i = 1, \ldots, n$ if $r \neq s$. Hence in every column of Π there is at most one single entry different from 0. But the entries of each column have to add up to 1, hence there is exactly one $t_s \in \{1, \ldots, m\}$ for each $s \leq n$ such that

$$\pi_{rs} = \delta_{rt_s}\,.$$

□

This theorem is rather remarkable. Under the included restrictions it tells us that if one performs a soft-aggregation, i. e. a not deterministic coarse-graining, with a reversible upper process ψ, the upper and lower processes cannot commute at all, and also the information flow does not vanish due to theorem 4.1. But being reversible for a process is - at least from the mathematical point of view - a rather weak restriction because the invertible matrices form an open and dense subset in $M(n \times n; \mathbb{R})$. Therefore, it is almost sure that the reversibility condition is fulfilled.
But reversibility of the upper process really turns out to be necessary as the following example shows.

Example 4.1. Lemma 4.4 turns out to be wrong if the matrix $\widehat{P}$ of the upper process is not required to be invertible. Let us consider a Markov process on a state space $\{x_1, x_2, x_3, x_4\}$ with four elements and transisition kernel

$$P = \frac{1}{6} \begin{pmatrix} 1 & 1 & 2 & 2 \\ 1 & 1 & 2 & 2 \\ 2 & 2 & 1 & 1 \\ 2 & 2 & 1 & 1 \end{pmatrix} .$$

Let $p = 1/4(1, 1, 1, 1)$ be the initial distribution. We have $Pp = p$. Furthermore we choose a soft-aggregation described by the matrix

$$\Pi = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1-\varepsilon & \varepsilon \\ 0 & 0 & \varepsilon & 1-\varepsilon \end{pmatrix} ,$$

with $0 < \varepsilon < 1$. Then it is an easy calculation that $\widehat{P}\Pi = \Pi P$, with $\widehat{P} = \Pi P \Pi^{\dagger}$, hence Eq. (4.1) is fulfilled, although the coarse-graining is not deterministic.

Let us consider another interpretation of theorem 4.4. If we are in the realm of this theorem, i. e. all entries of p are different from zero, the upper process is reversible and observationally commutes with the lower one, then the fact that the coarse-graining is deterministic provides us with the insight that not only observational commutativity holds, but also the information flow vanishes, as corollary 4.2 tells us. But this implication also holds true if one drops the assumption of an initial distribution vector p with all entries different from zero.

Corollary 4.5. *Let equation Eq. (4.1) holds true, and the upper process be reversible. Then also Eq. (3.1) holds.*

Proof. Up to relabeling of the states, we can assume that the initial distribution vector has the form $p = (p_1, \ldots, p_r, 0, \ldots, 0)$, with $1 \leq r \leq n$ and $p_i \neq 0$. Then, due to lemma 4.3 observational commutativity Eq. (4.1) is equivalent to the commutation relation $\widehat{P}\tilde{\Pi} = \tilde{\Pi}\tilde{P}$ where we adopt the notation of lemma 4.3. Since the upper process is reversible we are in the situation of theorem 4.4, with $\tilde{\Pi}$ instead of Π, and $\tilde{P}$ instead of P. Hence $\tilde{\Pi}$ is deterministic. By an analogous argument as in the proof of corollary 4.2 the result follows. □

5. Commutativity - existence of a transition kernel

The method of a level identification which we discuss in this section was proposed by Görnerup and Jacobi [8].

Definition 5.1. We speak of a *commutativity*, if there exists an $m \times m$ matrix $\tilde{P}$ such that

$$\tilde{P}\Pi = \Pi P ,$$

with the matrices P and Π defined by Eq. (2.1) and Eq. (2.2).

From lemma 4.3 one immediately obtains

Corollary 5.1. *Let observational commutativity Eq. (4.1) hold true. Then there are matrices $\tilde{P}$, $\widehat{P}$ and $\tilde{\Pi}$ as in lemma 4.3 such that*

$$\widehat{P}\tilde{\Pi} = \tilde{\Pi}\tilde{P} ,$$

i. e. for the system reduced to the support of the prior p it exists a transition kernel, i. e. it commutes in the sense of definition 5.1.

Guided by lemma 4.3, it might be tempting to think that also the reverse is true. I. e. the existence of a transition kernel $\tilde{P}$, i. e. an $m \times m$ matrix, fulfilling the commutation relation $\tilde{P}\Pi = \Pi P$, guarantees also that observational commutativity Eq. (4.1) holds true. But there is a crucial difference in the definitions consisting of the subtle detail that in definition 5.1 one asks only for the existence of such a matrix $\tilde{P}$, or linear map respectively, such that the diagram

$$\begin{array}{ccc} \mathbb{R}^m & \overset{\tilde{P}}{\dashrightarrow} & \mathbb{R}^m \\ {\scriptstyle\pi}\uparrow & & \uparrow{\scriptstyle\pi} \\ \mathbb{R}^n & \xrightarrow{\quad P \quad} & \mathbb{R}^n \end{array}$$

FIGURE 2. Fig. (1) in the frame of linear algebra.

commutes. Definition 4.1 instead not only leads to the existence of such a matrix $\tilde{P}$ but also that it is of the form $\tilde{P} = \Pi P \Pi^\dagger$ - a harder condition. The equivalence of both notions is disproved by the following example.

Example 5.1. We consider a stationary Markov process on a state space with three elements, and a transition kernel given by the matrix

$$P = \frac{1}{8}\begin{pmatrix} 3 & 3 & 2 \\ 3 & 3 & 2 \\ 2 & 2 & 4 \end{pmatrix}.$$

We perform a *soft-aggregation* described by the matrix

$$\Pi^\varepsilon = \begin{pmatrix} 1-\varepsilon & 1-\varepsilon & \varepsilon \\ \varepsilon & \varepsilon & 1-\varepsilon \end{pmatrix}, \quad 0 \le \varepsilon \le 1\,.$$

Let us define

$$\tilde{P} = \frac{1}{4}\begin{pmatrix} 3-\varepsilon & 2-\varepsilon \\ 1+\varepsilon & 2+\varepsilon \end{pmatrix}.$$

We have

$$\begin{aligned} \tilde{P}\Pi^\varepsilon &= \frac{1}{4}\begin{pmatrix} 3-\varepsilon & 2-\varepsilon \\ 1+\varepsilon & 2+\varepsilon \end{pmatrix}\begin{pmatrix} 1-\varepsilon & 1-\varepsilon & \varepsilon \\ \varepsilon & \varepsilon & 1-\varepsilon \end{pmatrix} \\ &= \frac{1}{4}\begin{pmatrix} (1-\varepsilon)(3-\varepsilon)+\varepsilon(2-\varepsilon) & (1-\varepsilon)(3-\varepsilon)+\varepsilon(2-\varepsilon) & \varepsilon(3-\varepsilon)+(1-\varepsilon)(2-\varepsilon) \\ (1-\varepsilon)(1+\varepsilon)+\varepsilon(2+\varepsilon) & (1-\varepsilon)(1+\varepsilon)+\varepsilon(2+\varepsilon) & \varepsilon(1+\varepsilon)+(1-\varepsilon)(2+\varepsilon) \end{pmatrix} \\ &= \frac{1}{4}\begin{pmatrix} 3-4\varepsilon+\varepsilon^2+2\varepsilon-\varepsilon^2 & 3-4\varepsilon+\varepsilon^2+2\varepsilon-\varepsilon^2 & 3\varepsilon-\varepsilon^2+2-3\varepsilon+\varepsilon^2 \\ 1-\varepsilon^2+2\varepsilon+\varepsilon^2 & 1-\varepsilon^2+2\varepsilon+\varepsilon^2 & \varepsilon+\varepsilon^2+2-\varepsilon-\varepsilon^2 \end{pmatrix} \\ &= \frac{1}{4}\begin{pmatrix} 3-2\varepsilon & 3-2\varepsilon & 2 \\ 1+2\varepsilon & 1+2\varepsilon & 2 \end{pmatrix}, \end{aligned}$$

and

$$\begin{aligned} \Pi^\varepsilon P &= \frac{1}{8}\begin{pmatrix} 1-\varepsilon & 1-\varepsilon & \varepsilon \\ \varepsilon & \varepsilon & 1-\varepsilon \end{pmatrix}\begin{pmatrix} 3 & 3 & 2 \\ 3 & 3 & 2 \\ 2 & 2 & 4 \end{pmatrix} = \frac{1}{8}\begin{pmatrix} 6-4\varepsilon & 6-4\varepsilon & 4 \\ 2+4\varepsilon & 2+4\varepsilon & 4 \end{pmatrix} \\ &= \frac{1}{4}\begin{pmatrix} 3-2\varepsilon & 3-2\varepsilon & 2 \\ 1+2\varepsilon & 1+2\varepsilon & 2 \end{pmatrix}, \end{aligned} \tag{5.1}$$

i. e. $\tilde{P}\Pi^\varepsilon = \Pi^\varepsilon P$. But for $\varepsilon \neq 0, 1/2, 1$ Eq. (4.1) can not be fulfilled at all. This can be seen as follows. Let us assume that Eq. (4.1) holds for an initial, stationary distribution $p = (p_1, p_2, p_3)$. Since every single entry of the matrix P is different from zero, one obtains from lemma 4.3 that also $p_i \neq 0$ for $i = 1, 2, 3$. Let $\widehat{P} = \Pi P \Pi^\dagger$. Due to lemma 4.3, Eq. (4.1) implies the commutation relation $\widehat{P}\Pi = \Pi P$. Eq. (5.1) shows rank $\Pi P = 2$ if $\varepsilon \neq 1/2$, hence also rank $\widehat{P} = 2$ and therefore the reversibility of the upper process. From theorem 4.4 one deduces that Π must be deterministic, which is wrong for $\varepsilon \neq 0, 1$. Therefore Eq. (4.1) leads to a contradiction.
Furthermore, by a tedious computation, one gets

$$\widehat{P} = \tilde{P} + \frac{\varepsilon(1-\varepsilon)}{4}\begin{pmatrix} \frac{-1}{\widehat{p}_1} & \frac{1}{\widehat{p}_2} \\ \frac{1}{\widehat{p}_1} & \frac{-1}{\widehat{p}_2} \end{pmatrix},$$

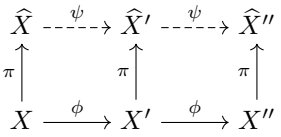


FIGURE 3. Same as Fig. (1), but expanded to show two consecutive time steps.

with $\widehat{p} = \Pi p$, and $\widehat{p} = (\widehat{p}_1, \widehat{p}_2)$. This formula indicates that Eq. (4.1) follows from the existence of a transition kernel $\tilde{P}$ if the aggregation is deterministic, i. e. if $\varepsilon = 0, 1$. This is indeed correct.

Proposition 5.2. *Let the aggregation map* Π *be deterministic. If it exists a transition kernel* $\tilde{P}$ *such that* $\tilde{P}\Pi = \Pi P$ *- i. e. the upper and lower processes commute - then Eq. (4.1) holds.*

Proof. For simplicity we assume that all entries p_r of the prior p are different from zero and that the aggregation map Π has full rank. This can be done without loss of generality since as in lemma 4.3, by relabeling the entries p_r of the prior and the corresponding columns and rows of the matrices involved, one can always arrive at a reduced system which fulfills these conditions.
Since Π is assumed to be deterministic, for all $r = 1, \ldots, n$ there is one $s_r \in \{1, \ldots, m\}$ such that $\pi_{sr} = \delta_{ss_r}$ where $\delta_{..}$ denotes the Kronecker-delta. This yields $\pi_{tr}\pi_{sr} = \delta_{ts_r}\delta_{ss_r} = \delta_{ts}\delta_{ss_r} = \delta_{ts}\pi_{sr}$ and we obtain

$$\begin{aligned}\sum_{r=1}^{n} \pi_{tr}\pi^{\dagger}_{rs} &= \frac{\sum_{r=1}^{n} \pi_{tr}\pi_{sr}p_r}{\widehat{p}_s} \\ &= \frac{\sum_{r=1}^{n} \pi_{tr}\pi_{sr}p_r}{\sum_{r=1}^{n} \pi_{sr}p_r} \\ &= \delta_{ts}\frac{\sum_{r=1}^{n} \pi_{sr}p_r}{\sum_{r=1}^{n} \pi_{sr}p_r} \\ &= \delta_{ts}\,,\end{aligned}$$

for all $t, s = 1, \ldots, m$. In the last step we made use of the assumption that the entries p_r of the prior are all different from zero, and that the aggregation map Π has full rank, because otherwise the fraction beside the δ_{ts} term can be equal zero.
The previous computation proves $\Pi\Pi^{\dagger} = 1_{m\times m}$, i. e. $\Pi^{\dagger}$ is a pseudo-inverse of Π. This yields

$$\tilde{P} = \tilde{P}1_{m\times m} = \tilde{P}\Pi\Pi^{\dagger} = \Pi P\Pi^{\dagger}$$

and the proof is done. □

6. MARKOVIANITY

In order to investigate how the already presented ideas of level identification relates to the fourth idea of Markovianity we need to extend the diagram by one consecutive time step (see Fig. (3)).

Definition 6.1. The observable process $\widehat{X}, \widehat{X}', \widehat{X}'', \ldots$ is itself a Markov process iff

$$I(\widehat{X} : \widehat{X}''|\widehat{X}') = 0\,, \tag{6.1}$$

i.e. $\widehat{X}$ and $\widehat{X}''$ are independent given $\widehat{X}'$.

We can show that Eq. (6.1) holds whenever the process is informationally closed, i.e. Eq. (3.1) holds true.

Theorem 6.1. *Eq. (3.1) implies Eq. (6.1).*

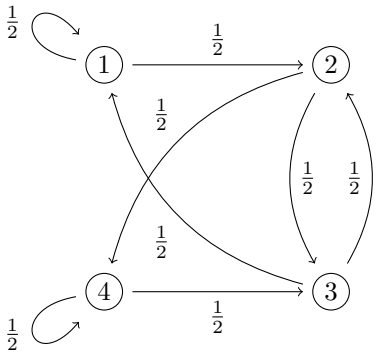
Proof. Consider the following mutual information

$$I(\widehat{X}'' : X'\widehat{X}|\widehat{X}') = \underbrace{I(\widehat{X}'' : X'|\widehat{X}')}_{=0 \text{ by Eq. (3.1)}} + \underbrace{I(\widehat{X}'' : \widehat{X}|X, \widehat{X}')}_{=0 \text{ according to Fig. (3)}} = 0$$

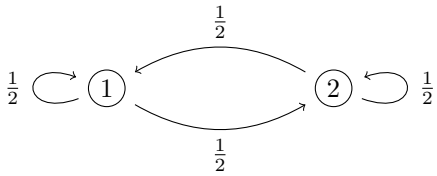
and the theorem follows from $I(\widehat{X}'' : \widehat{X}|\widehat{X}') \leq I(\widehat{X}'' : X\widehat{X}|\widehat{X}') = 0$. □

In general, from the markovianity Eq. (6.1) it is neither possible to infer that Eq. (3.1), Eq. (4.1) nor that commutativity, see definition 5.1, holds - even when the the coarse-graining is deterministic. An important example for systems that allow for Markovian coarse grained descriptions, but are not informationally closed with respect to a refined Markov partition are chaotic systems with Markov partitions.

Example 6.1. As stationary Markov process, we consider the Bernoulli-shift on a state space $\{x_1, x_2, x_3, x_4\}$ with four elements whose transition kernel can be illustrated as follows:



We perform a deterministic aggregation by lumping together the first with the second, and the third with the fourth state. This yields an i.i.d. - hence Markovian - process on a state space $\{\widehat{x}_1, \widehat{x}_2\}$



But Eq. (4.1) does not hold true, since

$$\begin{aligned} p(\widehat{X}' = \widehat{x}_1 | X = x_1) &= \sum_{l=1}^{4} p(\widehat{X}' = \widehat{x}_1 | X' = x_l) p(X' = x_l | X = x_1) \\ &= \quad p(\widehat{X}' = \widehat{x}_1 | X' = x_1) p(X' = x_1 | X = x_1) \\ &\quad + p(\widehat{X}' = \widehat{x}_1 | X' = x_2) p(X' = x_2 | X = x_1) \\ &= \frac{1}{2} + \frac{1}{2} = 1\,, \end{aligned}$$

but

$$\begin{aligned} p_{upper}(\widehat{X}' = \widehat{x}_1 | X = x_1) &= \sum_{r=1}^{2} p(\widehat{X}' = \widehat{x}_1 | \widehat{X} = \widehat{x}_r) p(\widehat{X} = \widehat{x}_r | X = x_1) \\ &= p(\widehat{X}' = \widehat{x}_1 | \widehat{X} = \widehat{x}_1) p(\widehat{X} = \widehat{x}_1 | X = x_1) \\ &= \frac{1}{2}\,. \end{aligned}$$

Hence also Eq. (3.1) can not hold, due theorem 4.1. Since the coarse-graining was deterministic, from proposition 5.2 follows that the upper and lower process do not commute.

7. Conclusions

On the one hand, the previous investigation poses several questions which can be divided into two categories. Firstly, some of the results achieved by our group in this paper have a rather formal character. This concerns in particular the difference between the two notions of observational commutativity and commutativity. Despite the fact we are able to give an example where these two ideas of a level identification turn out to be different, there is a lack of a precise interpretation of this result. Due to our understanding an independent description of the dynamics on the macrostates given by a transition kernel $\tilde{P}$ should also imply that Fig. (1) observationally commutes. And even though example 5.1 tells us that this implication is invalid, its rather formal character does not provide us with any clue about

the intrinsic reasons for this difference.
Additionally, there is also a lack of interpretation of the reversibility notion which is introduced in order to prove that observational commutativity even forces the aggregation to be deterministic. Formally, also in this case everything is clear: reversibility refers to the condition that $\widehat{P}$ is invertible. But it is part of future work to reformulate this condition in terms of stochastic processes.
The second category of open questions refers to the unproven implications. So far we could not prove equivalence of vanishing informational flow and observational commutativity in general. Also Markovianity is a rather isolated notion so far. It is the weakest of all we have introduced in this paper, but there are no proofs or disproofs whether observational commutativity or commutativity suffices to ensure a Markovian process on the upper level.

On the other hand, we achieved not only a thorough comparison between different closure measures, but also introduced two new ones formally: informational closure and observational commutativity. As far as we know, the present paper is the first attempt where the presented measures have been systematically proven to be different, at least if one considers soft-aggregations, whereas in the literature these measures have been announced to be equal. This holds in particular true for the last two closure notions: those of commutativity and Markovianity. At least in [8] both concepts were treated rather equivalently, whereas they turned out to be very different, even in the case of deterministic lumpings. Markovianity is a far weaker requirement a closed level description needs to fulfill than commutativity, as example 6.1 shows. The disagreement between the different notions is getting worse if one considers also soft-aggregations. In this case even the first three notions of level identification - informational closure, observational commutativity and commutativity - do not need to be equivalent at all.

8. Acknowlegement

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n° 267087 and n° 318723. Nils Bertschinger was supported by the Klaus Tschira Stiftung.

References

1. N. Ay and O. Pfante, *Finding closed sub-dynamics of dynamical systems*, in preparation.
2. Nils Bertschinger, Eckehard Olbrich, Nihat Ay, and Jürgen Jost, *Information and closure in systems theory*, Explorations in the Complexity of Possible Life. Proceedings of the 7th German Workshop of Artificial Life. (Amsterdam) (S. Artmann and P. Dittrich, eds.), IOS Press BV, 2006, pp. 9–21.
3. ———, *Autonomy: an information theoretic perspective.*, Biosystems **91** (2008), no. 2, 331–345.
4. C. K. Burke and M. Rosenblatt, *A markovian function of a markov chain*, Annals of Statistics **29** (1958), 1112–1122.
5. Thomas M. Cover and Joy A. Thomas, *Elements of information theory*, John Wiley & Sons, 2006.
6. Olof Görnerup and Martin Nilsson Jacobi, *A dual digenvector dondition for strong lumpability of markov chains*, SFI WORKING PAPER (2008), 1–7.
7. Olof Görnerup and Martin Nilsson Jacobi, *A method for inferring hierarchical dynamics in stochastic processes*, Adv.Complex Syst. **11** (2008), 1–16.
8. ———, *A method for finding aggregated representations of linear dynamical systems*, Adv.Complex Syst. **13** (2010), no. 2, 199–215.
9. N. Israeli and N Goldenfeld, *Coarse-graining of cellular automata, emergence, and the predictability of complex systems*, Physical Review E **73** (2006), 026203.
10. J. L. Snell J. G. Kemeny, *Finite markov chains*, Springer-Verlag, 1976.
11. D. V. Lindley, *Introduction to probability & statistics*, Cambridge University Press, 1965.
12. Cosma Rohilla Shalizi and Christopher Moore, *What is a macrostate? subjective observations and objective dynamics*, Tech. report, 2003, arXiv:cond-mat/0303625v1.

[1] Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22,
04103 Leipzig, Germany
[2] Santa Fe Insitute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA
E-mail address: [A)]pfante@mis.mpg.de, [B)]bertschi@mis.mpg.de, [C)]olbrich@mis.mpg.de, [D)]nay@mis.mpg.de, [E)]jost@mis.mpg.de