Max-Planck-Institut für Mathematik in den Naturwissenschaften Leipzig

The Information Bottleneck Method for Optimal Prediction of Multilevel Agent-based Systems

by

Robin Lamarche-Perrin, Sven Banisch, and Eckehard Olbrich

Preprint no.: 55 2015

The Information Bottleneck Method for Optimal Prediction of Multilevel Agent-based Systems

Preprint* of the Max Planck Institute for Mathematics in the Sciences

September 1st, 2015

Robin Lamarche-Perrin, Sven Banisch, and Eckehard Olbrich

Max Planck Institute for Mathematics in the Sciences
Inselstraße 22, 04103 Leipzig, Germany

Robin.Lamarche-Perrin@mis.mpg.de
Sven.Banisch@UniVerseCity.de
Eckehard.Olbrich@mis.mpg.de

Abstract

Because the dynamics of complex systems is the result of both decisive local events and reinforced global effects, the prediction of such systems could not do without a genuine multilevel approach. This paper proposes to found such an approach on information theory. Starting from a complete microscopic description of the system dynamics, we are looking for observables of the current state that allows to efficiently predict future observables. Using the framework of the Information Bottleneck method, we relate optimality to two aspects: the complexity and the predictive capacity of the retained measurement. Then, with a focus on Agent-based Models, we analyse the solution space of the resulting optimisation problem in a generic fashion. We show that, when dealing with a class of feasible measurements that are consistent with the agent structure, this solution space has interesting algebraic properties that can be exploited to efficiently solve the problem. We then present results of this general framework for the Voter Model with several topologies and show that, especially when predicting the state of some sub-part of the system, multilevel measurements turn out to be the optimal predictors.

Keywords: Information Theory, Information Bottleneck, Efficient Prediction, Multilevel Systems, Agent-based Models, Voter Model.

*This paper has been submitted in September 2015 to *Advances in Complex Systems*.

Contents

Table of Notations

$N \in \mathbb{N}$	number of agents
$\Omega = \{1, \ldots, N\}$	agent set
$A \subset \Omega$	agent subset
$\mathcal{A} = \{A_1, \ldots, A_k\} \subset 2^\Omega$	collection of agent subsets
S	agent state space
$\Sigma = S^N$	system state space
$\mathcal{S}_\phi$	state space of measurement ϕ
$X_i \in S$	state of agent i
$X = (X_1, \ldots, X_N) \in S^N$	system state
$X_A = (X_i)_{i \in A} \in S^{\lvert A \rvert}$	state of agents in subset A
$T(X^{t+1} \lvert X^t)$	Markov kernel
$t \in \mathbb{N}$	current time
$\tau \in \mathbb{N}$	prediction horizon
$X^0 \in \Sigma$	initial system state
$X^t \in \Sigma$	current system state
$X^{t+\tau} \in \Sigma$	future system state
$\phi : \Sigma \to \mathcal{S}_\phi$	pre-measurement (used for prediction)
$\psi : \Sigma \to \mathcal{S}_\psi$	post-measurement (to be predicted)
$\{\mu_A : \Sigma \to \mathcal{S}_\mu\}_{A \subset \Omega}$	generic measurement: a family of feasible measurements parametrized by an agent subset A
$\{\eta_A : \Sigma \to \mathbb{N}\}_{A \subset \Omega}$	the aggregated-state generic measurement
μ_A	feasible measurement: μ applied to agent subset A
$\mu_\mathcal{A} = (\mu_{A_1}, \ldots, \mu_{A_k})$	combination of feasible measurements: μ applied to a collection $\mathcal{A} = \{A_1, \ldots, A_k\}$ of agent subsets
$\mu_{\{i\}}$	agent measurement: μ applied to agent i
$(\mu_{\{1\}}, \ldots, \mu_{\{N\}})$	microscopic measurement: μ applied to each agent of the agent set Ω
$\mu_\Omega = \mu_{\{1,\ldots,N\}}$	macroscopic measurement: μ applied to the agent set Ω
$\mu_\emptyset$	empty measurement: no measurement is actually performed
$\beta \in \mathbb{R}^+$	trade-off parameter of the IB-variational
$IB_\beta(X; \hat{X}; Y) \in \mathbb{R}$	IB-variational
$\beta^{t,\tau}_{\phi_1,\phi_2} \in \mathbb{R}^+$	(if it exists) unique value of the trade-off parameter β for which the two pre-measurements ϕ_1 and ϕ_2 are IB-equivalent

1 Introduction

Because the dynamics of complex systems is the result of both decisive local events and reinforced global effects, the analysis, control and prediction of such systems could not do without a genuine multilevel approach. Typically complex systems can be observed at various scales and levels, but it is usually not obvious which of these observation levels is the most powerful to anticipate the system's dynamics. In the economy, for instance, one might rely on macro-economic observables like the GDP or its growth rate in order to anticipate the future performance of a country, but one might obtain a more differentiated picture by taking into account more refined information such as sectoral data, the current political situation, business climate indices or likewise information about the development of the most important trading partners. In weather forecast, the current atmospheric conditions are nowadays measured by a distributed web of weather stations harvesting micro-data regarding wind, temperature, humidity, *etc.* and this information is further complemented by macro-data from weather satellites and sounding balloons with the aim to make accurate predictions less expensive. How to decide how to combine such micro- and macro-data in an optimal fashion in order to provide a clear picture of future dynamics while avoiding the curse of complexity at the microscopic level? This paper addresses such questions related to multiscale measurements of complex systems in a theoretical setting. More precisely, it aims at emphasising the need for multilevel prediction in canonical examples of dynamical systems, and at founding such an approach on information theory.

Our general framework is the following: We start with a microscopic description of the system that is complete in the sense that it contains all available information about the system and its future. This means in particular that we assume the micro-dynamics to be Markovian. An example would be the phase space for classical particles consisting of their locations and velocities together with the Newtonian equations of motion. Another class of examples are models of interacting agents – also known as Agent-based Models (ABMs) – which typically implement a Markov chain on the state space defined by all possible agent configurations [5]. In this paper, we shall use a model of this latter type in order to present and elaborate the proposed prediction framework. Notice, that we consider only systems with discrete state spaces for this paper.

We are then interested in predicting a certain observable of the system which is determined by its future state via a stochastic map – called "post-measurement" in the following. In this paper, we will study both microscopic and macroscopic post-measurements. We are now asking for an observable of the current state – via a stochastic map called "pre-measurement" – which allow to most efficiently predict the target post-measurement. In this context, optimality is related to two aspects of the pre-measurement: (1) its predictive power and (2) its complexity. If one is only interested in predictive power the answer is trivial: The microscopic state would be the best pre-measurement because we assumed it being "complete". However, if complexity is also an issue one has to deal with a trade-off between these two competing aspects: How much predictive power does one is willing to

lose in order to reduce the cost of measurement? We address this question in the framework of the Information Bottleneck (IB) method [30]. Here, predictive power is quantified by the mutual information between the pre- and the post-measurement, while complexity is quantified by the mutual information between the pre-measurement and the current micro-state. This choice hence relates the complexity "cost" directly to the channel capacity that is required between the micro-state and the pre-measurement. Then, we define the Optimal Prediction Problem (OPP) as a constraint optimisation problem aiming at optimising a trade-off between measurement complexity and predictive capacity. The solution space is hence the set of all possible pre-measurement of the current system state.

Note that this choice of "model costs" does not take into account the difficulty of estimating the model from data. This aspect is not relevant for the current paper because we start here with the microscopic dynamics being known. Moreover, we also do not take into account neither the computational complexity of the prediction algorithm nor the real world costs of actually doing the "pre-measurements", *i.e.* the cost of data acquisition. The trade-off between prediction accuracy and computational costs is also addressed in the State Space Compression (SSC) framework developed in [31]. In particular several possibilities for information theoretic cost functions for both aspects are discussed. A principle difference between the SSC framework and the OPP formulated above is that the latter only asks for the mutual information between the pre-measurement and the post-measurement, but does not consider the construction of an explicit predictor, for instance by using an approximate dynamics for the pre-measurement. The latter is done in the SSC which in contrast to the OPP requires to consider also the computational costs for iterating this dynamics. The IB method applied to questions of prediction efficiency is clearly related to other information-theoretic approaches developed in the context of multilevel complex systems, and most importantly to predictive efficiency, introduced in [28] to characterise emergent levels. The intuition of this approach is that a coarse-grained description should be considered as a proper observational level if it informs about the dynamics that can be observed within this scale while, at the same time, being not too complex. Shalizi [28] introduced predictive efficiency as the ratio between excess entropy and statistical complexity [11], also known as effective measure complexity [16] and forecasting complexity [32]. The formulation as a variational of one-step mutual (prediction) information and the entropy of the description used in [24] renders the relation between predictive efficiency and the Information Bottleneck method visible.
There is also a strong relation of the OPP to the problem of level identification as for instance discussed in [25]. For level identification one asks for aggregations of the micro-state that give rise to a self-sufficient description on the aggregated level. In our formalism this mean that the post-measurement is pre-defined externally but varied together with the pre-measurement. In the context of Markov chains the existence of closed aggregated descriptions is denoted as "lumpability" of the Markov chain [18].

Our approach also relies on another essential feature of efficient prediction, that is, when dealing with a particular dynamical system, not all measurements are

meaningful and/or feasible in practice. Indeed, when observing a system, one might have at her disposal a determined collection of observation devices that one would like to use for optimal prediction: *e.g.*, a thermometer to measure the temperature of a gas, weather stations to get data about atmospheric conditions, or some well-defined economic indicators to estimate the performance of countries. The OPP should fit with these practical constraints to find, given such a collection of devices, the optimal way to position them within the system in order to retrieve informative data. Therefore, in this paper, we propose to constraint the solution space of the OPP by expressing the classes of pre-measurements that are both meaningful for the observer and feasible in practice. To focus on ABMs, we propose the concept of "generic measurement", that is an observation device that one can generically apply to any subset of agents in order to observe the system at different levels, from the agent microscopic level to the system macroscopic level through any intermediate mesoscopic level. A generic measurement hence defines a structured family of multilevel feasible measurements that can be used as a solution space for the OPP. Moreover, we show that such families are partially ordered by a so called "refinement relation" and that the IB-measures are monotonous with respect to this partial order, such that one always reduce the complexity and predictive capacity by moving upward within the solution space. Generic measurements hence prove to be extremely useful to explore the solutions of the optimisation problem.

A main result of the present investigation is that, depending on the weight of the measurement complexity in the IB trade-off, the optimal pre-measurement can be macroscopic, mesoscopic, microscopic, or multilevel, *i.e.* a combination of observables from different levels. In ABMs, for instance, we might ask (and we will in Section 4) what is the most efficient measurement that one can perform on the current system state (at time t) in order to predict observables of the future system state (at time $t+\tau$) that can be (1) global observables of the whole agent population, (2) observables focusing on a subset of this population, or just as well (3) observables of the state of one particular agent of interest. The aim of the framework we propose is to identify pre-measurements that are optimal in the sense that they provide the best predictive power at a certain level of allowed complexity. To make this illustration a bit more explicit, for the prediction of the state of an individual agent at time $t+\tau$, knowledge of its state at time t is often the best choice for short-term prediction (small τ). Indeed, this local observable is of rather low complexity, since the number of possible values is equal to the number of states the agent may adopt, while the predictive information concerning the future state of this agent is relatively high in every ABM in which agents are updated sequentially so that only a fraction of agents changes from one time step to the next. As the prediction horizon increases (larger τ), however, other measurements – capturing mesoscopic or macroscopic information, or even combinations of both – become optimal, since for longer timescales the dynamics are governed by processes on larger spacial scales.

These different effects are already visible in the Voter Model (VM) [7,8,19,21,23] that we use as a paradigmatic ABM example, simple enough to compute the involved information measures in an explicit way on the basis of the microscopic

transition matrices. For this model, we define and compute "IB-diagrams" showing pre-measurements that are optimal according to the IB objective function within regions of a three-parameter space, spanned by the current observation time t, the prediction horizon τ, and the trade-off parameter β between measurement complexity and predictive capacity. We present results for the VM on the complete graph, on a two-community graph, and on the ring. These three experiments show that, depending on the prediction horizon, higher-level and even multilevel measurements turn out to be the optimal predictors maximising information for a given complexity.

The rest of this paper is organised as follows. Section 2 presents the Information Bottleneck (IB) method, the Optimal Prediction Problem (OPP), the concept of "generic measurement", and its application to Agent-based Models (ABMs). Section 3 presents general results regarding the solution space of the OPP: poset structure of feasible measurements, monotonicity of IB measures, optimality conditions for some subsets of the solution space. Section 4 applies our framework to the Voter Model (VM) and provides a complete solution of the OPP in a very simple case: predicting the macroscopic "aggregated-state" measurement in the case of a complete and uniform interaction graph. Section 5 presents more complex experiments where a multilevel measurement becomes necessary for optimal prediction. We show how by introducing heterogeneity in the interaction graph of the VM affects the prediction problem by creating a dependence between the system's temporality and the optimal levels for prediction. Section 6 discusses these results and proposes some application and generalisation perspectives.

2 Applying the Information Bottleneck Method to Optimal Prediction of Agent-based Models

First, this section formalises the Optimal Prediction Problem (OPP) by applying the Information Bottleneck (IB) method to the measurement and prediction of Markov chains (Subsection 2.1). We then define *IB-diagrams* as solutions to the OPP, partitioning the tridimensional parameter space (current time, prediction horizon, and trade-off parameter) into disjoint *optimality regions* for each possible pre-measurements (Subsection 2.2). We apply this general framework to the prediction of ABMs by defining the concept of *generic measurement*, that is a family of pre-measurements that are consistent with the agent structure of the system (Subsection 2.3). More precisely, they are consistent in the sense that such a generic measurement can be virtually applied to any subset of agents while satisfying two essential algebraic properties : independence of measurements relatively to the state of non-observed agents, and additivity of observables of disjoint agent subsets. Section 3 will then show that, under these conditions, the set of feasible pre-measurements has an interesting poset structure that is consistent with the IB framework.

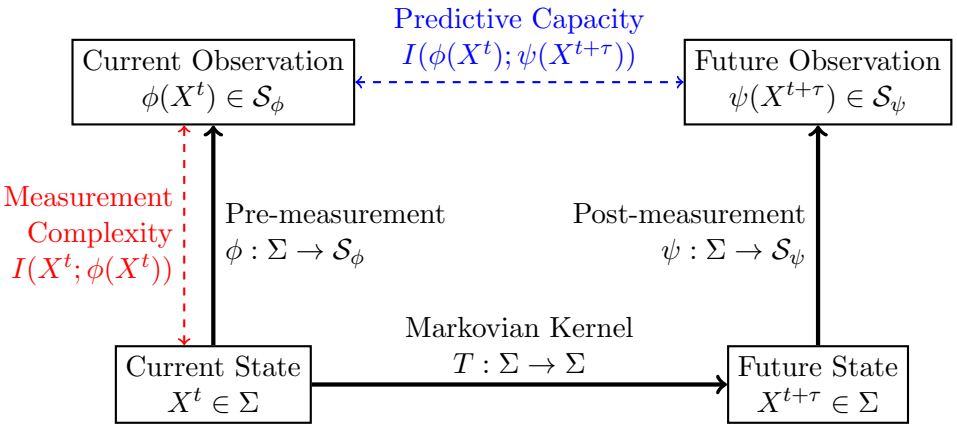


Figure 1: General setting of the Optimal Prediction Problem

2.1 Information Bottleneck and Optimal Prediction

Let $(X^t)_{t\in\mathbb{N}}$ be a sequence of discrete random variables defined on a finite state space Σ and representing the trajectory of a dynamical system assumed to have the Markov property. We mark $T(X^{t+1} = x'|X^t = x)$ the transition probability at time t. A *measurement* ϕ is a stochastic map that transforms $X \in \Sigma$ into a random variable $\phi(X) \in \mathcal{S}_\phi$ defined on another finite state space $\mathcal{S}_\phi$ – called the *measurement space* – according to a conditional probability distribution $\Pr(\phi(X) = s_\phi|X = x)$. A measurement hence induces a *soft partitioning* $s_\phi \in \mathcal{S}$ of the system state space $x \in \Sigma$. We assume in the following that both the transition kernel and the measurements are time-independent. Now, given a *post-measurement* ψ that will be performed at time $t+\tau$, we would like to perform a *pre-measurement* ϕ at time t that efficiently predicts the result of the future measurement (see Fig. 1 for a graphical representation of this general setting).

The IB-method [30] is an information-theoretic framework generalising rate distortion theory [10] to build lossy statistics while controlling the information loss. It consists in solving the following optimisation problem: given two random variables X and Y – respectively the *input* and the *output* of the IB method, find a third *bottleneck* variable $\hat{X}$ that is a (possibly stochastic) function of X extracting and compressing the information its contains that is relevant to predict Y. The bottleneck variable $\hat{X}$ is thus a model of X that is adequate to explain Y. In our setting the input X corresponds to the current state X^t, the output Y to the post-measurement $\psi(X^{t+\tau})$ and the bottleneck $\hat{X}$ to the pre-measurement $\phi(X^t)$. Interpreting the bottleneck variable as a model of the output, the *complexity* of this model is quantified by the mutual information $I(X^t; \phi(X^t))$, that is the amount of information that the compressed variable contains about the data it summarises. The *predictive capacity* of the model is quantified by $I(\phi(X^t); \psi(X^{t+\tau}))$, that is the amount of information captured by $\phi(X^t)$ that can be used to predict $\psi(X^{t+\tau})$.

The IB method is thus actually dealing with a constrained optimisation problem: *Build a bottleneck variable that optimally predicts the output without exceeding a given complexity*; or its dual version: *Build a bottleneck variable with minimal complexity that guarantees a given predictive capacity.* Usually, both conditions are expressed by the following variational problem [30]:

$$\min_{\phi} IB_\beta(X^t; \phi(X^t); \psi(X^{t+\tau}))$$

- where ϕ is a stochastic map from Σ to any discrete space $\mathcal{S}_\phi$,
- where $IB_\beta(X^t; \phi(X^t); \psi(X^{t+\tau})) \;=\; I(X^t; \phi(X^t)) \;-\; \beta\, I(\phi(X^t); \psi(X^{t+\tau}))$,
- and where $\beta \in \mathbb{R}^+$ expresses the trade-off between the two competing goals.

As $\beta \to 0$, one focuses on compression against prediction whereas, as $\beta \to +\infty$, predictive capacity prevails. In this paper, we are hence interested in the following class of optimisation problems.

Definition 1 (Optimal Prediction Problem)**.** *Given an initial distribution* $\Pr(X^0 = x)$*, a transition kernel* $T(X^{t+1} = x' | X^t = x)$*, and a post-measurement* ψ *defined by* $Pr(\psi(X) = s_\psi | X = x)$*, an* Optimal Prediction Problem *(OPP) is an instance of the following optimisation problem:*

- *Given a time* $t \in \mathbb{N}$*, an horizon* $\tau \in \mathbb{N}$*, and trade-off parameter* $\beta \in \mathbb{R}^+$*,*
- *Find a pre-measurement* ϕ *defined by* $\Pr(\phi(X) = s_\phi | X = x)$ *that minimises the IB-variational* $IB_\beta(X^t; \phi(X^t); \psi(X^{t+\tau}))$*.*

2.2 Computing Information Bottleneck Diagrams

An instance of a prediction problem – as hereabove defined – is characterised by three parameters: the current time of pre-measurement $t \in \mathbb{N}$, the prediction horizon before post-measurement $\tau \in \mathbb{N}$, and the trade-off parameter $\beta \in \mathbb{R}^+$. Hence, for a given dynamical system and a given post-measurement ψ, a complete solution to the OPP consists in finding an optimal measurement for each triple $(t, \tau, \beta) \in \mathbb{N} \times \mathbb{N} \times \mathbb{R}^+$ or, conversely, in identifying the *optimality region* of each possible pre-measurement ϕ, that is the subset of $\mathbb{N} \times \mathbb{N} \times \mathbb{R}^+$ where ϕ is optimal. We call the resulting partitioning of the tridimensional parameter space an *IB-diagram*.

To build and analyse such diagrams, we actually look for the *borders* between optimality regions, that is the values of (t, τ, β) where the IB-variational is equal for two given pre-measurements ϕ_1 and ϕ_2, thus delimiting two regions of $\mathbb{N} \times \mathbb{N} \times \mathbb{R}^+$, one where ϕ_1 is more efficient than ϕ_2, and the other where ϕ_2 is more efficient than ϕ_1. Given a finite set of pre-measurements $\{\phi_1, \ldots, \phi_k\}$, and assuming that one can easily determine such border for each pair (ϕ_i, ϕ_j), then one can deduce, for any triple (t, τ, β), which of these pre-measurements is optimal (by pairwise comparison).

The β-border is...	When...	Optimality region of ϕ_1	Optimality region of ϕ_2
(1) strictly positive	$H_1 < H_2$ and $I_1 < I_2$	$[0, \beta^{t,\tau}_{\phi_1,\phi_2}]$	$[\beta^{t,\tau}_{\phi_1,\phi_2}, +\infty[$
	$H_1 > H_2$ and $I_1 > I_2$	$[\beta^{t,\tau}_{\phi_1,\phi_2}, +\infty[$	$[0, \beta^{t,\tau}_{\phi_1,\phi_2}]$
(2) null	$H_1 = H_2$ and $I_1 < I_2$	$\{0\}$	$[0, +\infty[$
	$H_1 = H_2$ and $I_1 > I_2$	$[0, +\infty[$	$\{0\}$
(3) infinite	$H_1 < H_2$ and $I_1 = I_2$	$[0, +\infty[$	$\{+\infty\}$
	$H_1 > H_2$ and $I_1 = I_2$	$\{+\infty\}$	$[0, +\infty[$
(4) defined nowhere	$H_1 < H_2$ and $I_1 > I_2$	$[0, +\infty[$	$\emptyset$
	$H_1 > H_2$ and $I_1 < I_2$	$\emptyset$	$[0, +\infty[$
(5) defined everywhere	$H_1 = H_2$ and $I_1 = I_2$	$[0, +\infty[$	$[0, +\infty[$

where $H_i = I(X^t; \phi_i(X^t))$ is the complexity of ϕ_i,
$I_i = I(\phi_i(X^t); \psi(X^{t+\tau}))$ is the predictive capacity of ϕ_i,
$\beta^{t,\tau}_{\phi_1,\phi_2} = \frac{H_2 - H_1}{I_2 - I_1}$ is the strictly positive border between ϕ_1 and ϕ_2 in case (1).

Table 1: Characterising the optimality regions of two pre-measurements ϕ_1 and ϕ_2 along the dimension of the trade-off parameter β in the IB-variational

Table 1 provides an exhaustive list of the five possible types of β-borders – that is the values of the trade-off parameters where ϕ_1 and ϕ_2 are IB-equivalent, for a fixed $t \in \mathbb{N}$ and a fixed $\tau \in \mathbb{N}$ – depending on the complexity and the predictive capacity of the two competing measurements ϕ_1 and ϕ_2. In the following, we are essentially interested in case (1), that is when, for a fixed $t \in \mathbb{N}$ and a fixed $\tau \in \mathbb{N}$, there is a unique and strictly positive value $\beta > 0$ of the trade-off parameter for which the two measurements are IB-equivalent:

$$\exists! \beta \in]0, +\infty[, \quad IB_\beta(X^t; \phi_1(X^t); \psi(X^{t+\tau})) \;=\; IB_\beta(X^t; \phi_2(X^t); \psi(X^{t+\tau})).$$

In this case, we say that the border between the two optimality regions is *strictly positive*, and we mark $\beta^{t,\tau}_{\phi_1,\phi_2} \in]0, +\infty[$ the corresponding value, delimiting the two optimality regions. It is easy to show that this case arises if and only if the complexity and the predictive capacity of one measurement are both strictly larger than those of the other measurement, and that

$$\beta^{t,\tau}_{\phi_1,\phi_2} \;=\; \frac{I(X^t; \phi_2(X^t)) - I(X^t; \phi_1(X^t))}{I(\phi_1(X^t); \psi(X^{t+\tau})) - I(\phi_2(X^t); \psi(X^{t+\tau}))}.$$

In the experiments of Sections 4 and 5, we use this formula to compute and analyse

several such IB-diagrams.

For other cases, one measurement is never less efficient than the other one, meaning that one can safely chose this measurement for any value of the trade-off parameter. In this case, the two measurements can however be equivalent for extreme values of β (cases (2) and (3)) or for all values of β (case (5), which only arises when ϕ_1 and ϕ_2 have the same complexity and the same predictive capacity).

Remark on the Optimality Region of Deterministic Measurements. Since the complexity of any deterministic measurement ϕ_d is actually equal to its entropy:

$$I(X^t; \phi_d(X^t)) \quad = \quad H(\phi_d(X^t)) - H(\phi_d(X^t)|X^t) \quad = \quad H(\phi_d(X^t)),$$

and since the entropy is an upper bound of the predictive capacity:

$$I(\phi_d(X^t); \psi(X^{t+\tau})) \quad \leq \quad H(\phi_d(X^t)),$$

then the predictive capacity of any deterministic measurement is always lower than its complexity:

$$H(\phi_d(X)|X) = 0 \qquad \Rightarrow \qquad I(\phi_d(X^t); \psi(X^{t+\tau})) \quad \leq \quad I(X^t; \phi_d(X^t)).$$

Hence, $\forall \beta < 1,\ IB_\beta(X^t; \phi_d(X^t); \psi(X^{t+\tau})) > 0$.

Moreover, any "constant measurement" $\phi_c(X) = c \in \mathcal{S}_\phi$ has a null complexity and null predictive capacity: $\forall \beta \in \mathbb{R}^+,\ IB_\beta(X^t; \phi_c(X^t); \psi(X^{t+\tau})) = 0$. Hence, such a constant measurement is always more efficient than any deterministic measurements for $\beta < 1$.

2.3 Application to Agent-based Models via the Concept of Generic Measurement

Seeking for a pre-measurement that optimises the IB-variational might result in any kind of stochastic map. However, when measuring and predicting a dynamical process modelling a particular physical system, such an optimal solution might appear quite abstract or quite artificial in practice. In fact, one may not be able to physically implement the corresponding observation device. Hence, the solution space of the OPP, that is the set of all possible stochastic maps, should be redefined in order to fit with the practical measurement feasibilities. Given a set $\Phi = \{\phi_1, \ldots, \phi_k\}$ of measurements that are actually *feasible* in practice, we would like to find the one that optimises the IB trade-off:

$$\min_{\phi \in \Phi} IB_\beta(X^t; \phi(X^t); \psi(X^{t+\tau})).$$

We now consider the case of ABMs. Given a set $\Omega = \{1, \ldots, N\}$ of N agents indexed by integers, the state of agent i at time t is represented by a discrete random

variable $X_i^t \in S$, where S is the agent state space[1]. The state of the whole system is hence represented by the multivariate random variable $X^t = (X_1^t, \ldots, X_N^t)$ defined on the Cartesian product of the agent state spaces $\Sigma = S^N$. In the following, we introduce the concept of *generic measurement* to express the feasibility of pre-measurements as their consistency with this agent dimension.

Generic Measurement. Intuitively, a *generic measurement* is an observation procedure that can be independently applied to any agent subset. It thus defines a family of feasible measurements parametrized within the power set 2^Ω of the agent set Ω. For example, any statistic – that is any measure of some attribute of the agents (mean, variance, maximum, *etc.*) – is a generic measurement in the sense that it is a standard aggregating operation that one can generically apply to any sample of agents. For more applied examples, the sum of the incomes of individuals is a generic measurement in economy and, if one identifies gas particles with agents, then the mass, kinetic energy, and other classical physical properties of particles are also generic measurements. Note that most of these examples are also additive measurements, as defined further in this section.

Definition 2 (Generic Measurement)**.** *A* generic measurement μ *is a family of feasible measurements* $(\mu_A : \Sigma \to \mathcal{S}_\mu)_{A \subset \Omega}$ *parametrized by an agent subset* $A \subset \Omega$ *that all take values into a common measurement space* $\mathcal{S}_\mu$ *and that each is a (possibly stochastic) functions of the states of the agents in* A only*:*
$\forall A \subset \Omega,\ \forall x = (x_1, \ldots, x_N) \in S^N,\ \forall s_\mu \in \mathcal{S}_\mu,$

$$\Pr(\mu_A(X) = s_\mu \mid X = (x_1, \ldots, x_N)) \ = \ \Pr(\mu_A(X) = s_\mu \mid (X_i = x_i)_{i \in A}).$$

For any collection of agent subsets $\{A_1, \ldots, A_k\} \subset 2^\Omega$*, a generic measurement also defines a feasible measurement consisting in the combination of several measurements* $(\mu_{A_1}, \ldots, \mu_{A_k})$ *and taking values into the* k*-dimensional measurement space* $(\mathcal{S}_\mu)^k$*. In particular, a generic measurement* μ *defines (see Fig. 2 for a graphical representation of the following measurements):*

- *a set of* agent measurements $\mu_{\{i\}} : \Sigma \to \mathcal{S}_\mu$*, with* $i \in \Omega$*, each corresponding to the measurement of the state of a single agent;*
- *a* microscopic measurement $(\mu_{\{1\}}, \ldots, \mu_{\{N\}}) : \Sigma \to (\mathcal{S}_\mu)^N$ *corresponding to the combination of all agent measurements;*
- *a* macroscopic measurement $\mu_\Omega : \Sigma \to \mathcal{S}_\mu$ *corresponding to the measurement of the macroscopic state of the whole population;*
- *an* empty measurement $\mu_\emptyset : \Sigma \to \mathcal{S}_\mu$ *for which no observation is actually performed (*i.e.*, this measurement is constant in* $\mathcal{S}_\mu$*).*

[1] In this setting, we assume that all agents have a common state space S, thus assuming that they are somewhat homogeneous. However, one can always come down to this case by modelling the common state space S as the union $S_1 \cup \ldots \cup S_N$ of the particular state spaces for each agent in Ω (possibly inducing a very sparse transition matrix for the dynamical system).

As an example, consider agents accumulating and sharing a given resource such that the state of agent $i \in \Omega$ represents the current amount of resources it owns: $X_i \in S = \mathbb{N}$. A canonical generic measurement would consists in associating to each agent subset $A \subset \Omega$ the total amount of resources owned within the subset: $\mu_A(X) = \sum_{i \in A} X_i \in \mathcal{S}_\mu = \mathbb{N}$. Within this setting:

- an agent measurement $\mu_{\{i\}}(X) = X_i$ simply gives the amount of resources owned by agent i;
- the microscopic measurement $(\mu_{\{1\}}, \ldots, \mu_{\{N\}})(X) = (X_1, \ldots, X_N)$ provides a complete description of the system state by specifying the amount of resources owned by each agent separately;
- the macroscopic measurement $\mu_\Omega(X) = \sum_{i \in \Omega} X_i$ gives the total amount of resources owned within the system;
- and the empty measurement $\mu_\emptyset(X) = 0$, as in any other setting, do not provide any information regarding the current system state.

More generally, given a generic measurement μ, a *feasible measurement* ϕ can be expressed with respect to a *feasible collection of agent subsets* $\{A_1, \ldots, A_k\} \subset 2^\Omega$ – that is a collection of agent subsets to which the generic measurement μ can be applied in practice. In other words, the feasibility constraints regarding measurements are now fully expressed in terms of the agent space. For example, in the case of agents spatially located in an environment, the environment's topology might impose practical constraints for measurement (a thermometer for example measures a local property of the system and is hence sensitive to its topology).

Additive Generic Measurements. Intuitively, a generic measurement is additive if, when measuring the state of groups of agents at a given level, one can directly derive the measures of the states of groups of agents appearing at higher-levels. For example, one can derive any mesoscopic measurement from the microscopic one. This important property implies that measurements can be somehow ordered according to the information they provide about the system, as formalised in next section.

Definition 3 (Additivity)**.** *A generic measurement μ is* additive *if the measurement of two disjoint agent subsets fully determine the measurement of their union, and the measurement of two nested agent subsets fully determine the measurement of the corresponding complement. Formally, $\forall A_1 \subset \Omega$, $\forall A_2 \subset \Omega$, we have:*

$$A_1 \cap A_2 = \emptyset \quad \Rightarrow \quad \begin{cases} H(\mu_{A_1 \cup A_2}(X) \mid \mu_{A_1}(X),\ \mu_{A_2}(X)) & = 0 \\ H(\mu_{A_1}(X) \mid \mu_{A_1 \cup A_2}(X),\ \mu_{A_2}(X)) & = 0. \end{cases}$$

This is the case in particular when, given a commutative, symmetric, and invertible operator $\star$ on the measurement space $\mathcal{S}_\mu$, this operator is preserved by the

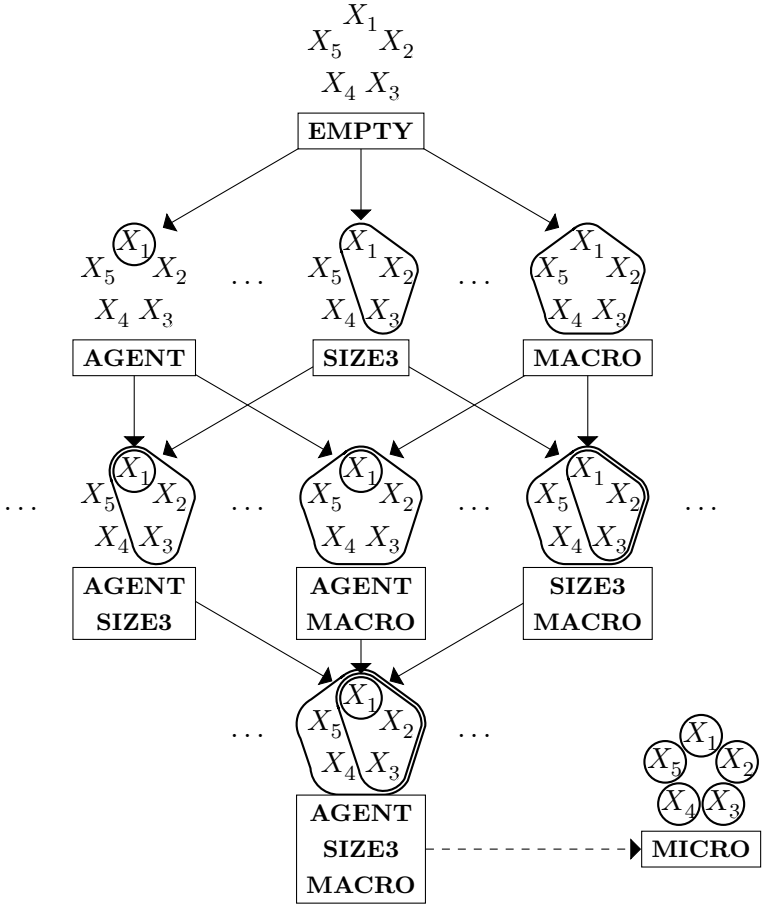


Figure 2: Poset of feasible measurements where arrows represent the refinement relation (in case of an additive generic measurement)

union of disjoint agent subsets. Formally, $\forall A_1 \subset \Omega$, $\forall A_2 \subset \Omega$, $\forall s_{\mu_1} \in \mathcal{S}_\mu$, and $\forall s_{\mu_2} \in \mathcal{S}_\mu$, we have:

$$A_1 \cap A_2 = \emptyset \;\Rightarrow\; \begin{cases} \Pr(\mu_{A_1 \cup A_2}(X) = s_{\mu_1} \star s_{\mu_2} \mid \mu_{A_1}(X) = s_{\mu_1},\ \mu_{A_2}(X) = s_{\mu_2}) & = 1 \\ \Pr(\mu_{A_1}(X) = s_{\mu_1} \mid \mu_{A_1 \cup A_2}(X) = s_{\mu_1} \star s_{\mu_2},\ \mu_{A_2}(X) = s_{\mu_2}) & = 1. \end{cases}$$

In this case, the probability distribution of $\mu_{A_1 \cup A_2}(X)$ is fully determined by the joint probability distribution of $\mu_{A_1}(X)$ and $\mu_{A_2}(X)$.

In our previous example regarding the amount of resources own by agents, with $\mathcal{S}_\mu = \mathbb{N}$, feasible measurements $\mu_A(X) = \sum_{i \in A} X_i$ are additive with respect to the addition $+$ in $\mathbb{N}$:

$$A_1 \cap A_2 = \emptyset \quad \Rightarrow \quad \mu_{A_1 \cup A_2}(X) = \mu_{A_1}(X) + \mu_{A_2}(X).$$

Higher-order Generic Measurements. The definition of generic measurements provided above is rather general and comprises a plethora of observables. There are, however, already relatively common cases for which this definition becomes problematic, because the measurement will not be additive. Examples would be observables that measure properties of pairs of agents such as the energy in the Ising Model or the number of active links in the Voter Model. The microscopic measurements as defined above would become meaningless in these cases. A simple solution for this problem is the definition of higher-order generic measurements that are defined not on subsets of agents but on a subsets of all pairs (2^{nd} order) – or in general k-tuples (k^{th} order) – of agents. The energy of an Ising system would then be an additive second-order generic measurement. In this paper, however, we only survey and use first-order generic measurements.

Deterministic Generic Measurements. To conclude this section, we also define deterministic generic measurements.

Definition 4 (Determinism)**.** *A generic measurement μ is* deterministic *if each feasible measurement μ_A with $A \subset \Omega$ is a deterministic function of the states of the agents in A* only*:*

$$H(\mu_A(X)|X_A) = 0,$$

where $X_A = (X_i)_{i\in A}$ is the microscopic state of agents in A.

Observation 1. *If μ is additive and deterministic, then the microscopic state can be used instead of the measurement in the additivity property. Formally, $\forall A_1 \subset \Omega$, $\forall A_2 \subset \Omega$, we have:*

$$A_1 \cap A_2 = \emptyset \quad \Rightarrow \quad \begin{cases} H(\mu_{A_1\cup A_2}(X) \mid \mu_{A_1}(X),\ X_{A_2}) & = 0 \\ H(\mu_{A_1}(X) \mid \mu_{A_1\cup A_2}(X),\ X_{A_2}) & = 0, \end{cases}$$

where $X_A = (X_i)_{i\in A}$ is the microscopic state of agents in A.

Proof. Since μ is deterministic, we have

$$\begin{aligned} &H(\mu_{A_1\cup A_2}(X)|\mu_{A_1}(X), \mu_{A_2}(X), X_{A_2}) \\ &\quad = H(\mu_{A_1\cup A_2}(X), \mu_{A_2}(X)|\mu_{A_1}(X), X_{A_2}) \ - \underbrace{H(\mu_{A_2}(X)|\mu_{A_1}(X), X_{A_2})}_{=0} \\ &\quad = H(\mu_{A_1\cup A_2}(X)|\mu_{A_1}(X), X_{A_2}) \ + \underbrace{H(\mu_{A_2}(X)|\mu_{A_1\cup A_2}(X), \mu_{A_1}(X), X_{A_2})}_{=0} \end{aligned}$$

and since μ is additive, we also have

$$H(\mu_{A_1\cup A_2}(X)|\mu_{A_1}(X), \mu_{A_2}(X), X_{A_2}) = 0.$$

The same reasoning also leads to $H(\mu_{A_1}(X)|\mu_{A_1\cup A_2}(X), X_{A_2}) = 0$. □

3 Theoretical Results regarding the Solution Space of the Optimal Prediction Problem

This section present some general properties of the solution space of the OPP – as defined in the previous section – that is the set of feasible measurements that can be derived from a generic measurement. First, we show that this solution space can be partially ordered by a "refinement relation" and that the complexity and the predictive capacity (see Subsection 2.1) are monotonous with respect to this poset structure. Second, we show that, under some additional assumptions, the solution space can be significantly reduced by *a priori* removing a subset of non-optimal measurements, thus reducing the computation cost of the optimisation problem.

3.1 The Poset of Feasible Measurements

Measurements can be partially ordered according to their relative information content. Intuitively, a measurement ϕ_1 "precedes" a measurement ϕ_2 if all the information contained in $\phi_2(X)$ regarding the system's state X is also contained in $\phi_1(X)$. In the case of deterministic measurements, this partial order corresponds to the classical *refinement relation* [12] between the two partitions of the state space induced by ϕ_1 and ϕ_2. In this context, a partition refines another partition if each part of the first partition is a subset of a part of the second partition. In the following definition, we keep the name of this partial order relation while generalising to any (stochastic or deterministic) measurement. Note that this generalised relation is more commonly known as the *Blackwell order* [6].

Definition 5 (Refinement Relation)**.** *A measurement ϕ_1* refines *a measurement ϕ_2 (we mark $\phi_1 \prec \phi_2$) if and only if $X \to \phi_1(X) \to \phi_2(X)$ is a Markov chain for any random variable X in Σ. In other words, ϕ_1 refines ϕ_2 if and only if $\phi_2(X)$ is a (possibly stochastic) function of $\phi_1(X)$* only*.*

For any time $t \in \mathbb{N}$ and horizon $\tau \in \mathbb{N}$, we hence have the following Bayesian network [14]:

$$\begin{array}{ccccc} X^t & \longrightarrow & \phi_1(X^t) & \longrightarrow & \phi_2(X^t) \\ \downarrow & & & & \\ X^{t+\tau} & \longrightarrow & \psi(X^{t+\tau}) & & \end{array}$$

Any collection of measurements hence forms a poset that can be represented by a Hasse diagram (see the example in Section 4). Moreover, this poset structure is consistent with the IB-measures as stated in the following theorem.

Theorem 1. *The measurement complexity and the predictive capacity are monotonous regarding the refinement relation:*

$$\phi_1 \prec \phi_2 \quad \Rightarrow \quad \begin{cases} I(X^t;\phi_1(X^t)) & \geq & I(X^t;\phi_2(X^t)) \\ I(\phi_1(X^t);\psi(X^{t+\tau})) & \geq & I(\phi_2(X^t);\psi(X^{t+\tau})) \end{cases}$$

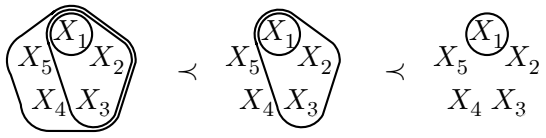
For any μ, 

Figure 3: Combining measurements generates refinements (see Observation 2)

Proof. These two inequalities are obtained by exploiting the conditional independences implied by the Bayesian network presented in Definition. 5. First, $\phi_2(X^t)$ is conditionally independent from X^t given $\phi_1(X^t)$, because of the Markov chain $X^t \to \phi_1(X^t) \to \phi_2(X^t)$. Due to this conditional independence the conditional mutual information vanishes:

$$I(X^t; \phi_2(X^t)|\phi_1(X^t)) = 0.$$

Now we can apply the chain rule to the following mutual information:

$$\begin{aligned} I(X^t; \phi_1(X^t), \phi_2(X^t)) &= I(X^t; \phi_1(X^t)) + \underbrace{I(X^t; \phi_2(X^t)|\phi_1(X^t))}_{=0} \\ &= I(X^t; \phi_2(X^t)) + I(X^t; \phi_1(X^t)|\phi_2(X^t)). \end{aligned}$$

Thus,

$$I(X^t; \phi_1(X^t)) - I(X^t; \phi_2(X^t)) = I(X^t; \phi_1(X^t)|\phi_2(X^t)) \geq 0,$$

which proves the upper inequality.

Second, $\psi(X^{t+\tau})$ is conditionally independent from $\phi_2(X^t)$ given $\phi_1(X^t)$, because $\phi_1(X^t)$ d-separates [14] $\psi(X^{t+\tau})$ and $\phi_2(X^t)$. Using the chain rule for $I(\psi(X^{t+\tau}); \phi_1(X^t), \phi_2(X^t))$ analogously to the first case proves the second inequality. Note that these inequalities are the Data Processing Inequalities [9] of the corresponding Markov chains. □

In the case of (possibly additive) generic measurement of agent-based systems, the following observations also apply (see Fig. 2, 3, 4 and 5).

Observation 2. *For any generic measurement μ, combining two feasible measurements generates a refining measurement:*

$$\forall A_1 \subset \Omega, \quad \forall A_2 \subset \Omega, \quad (\mu_{A_1}, \mu_{A_2}) \prec \mu_{A_1}.$$

Proof. Indeed, $(\mu_{A_1}, \mu_{A_2})(X)$ contains all the information needed to determine $\mu_{A_1}(X)$. □

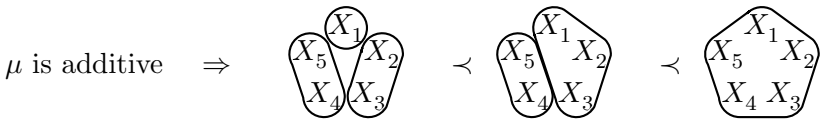


Figure 4: When additive, measurements of partitions refines measurements of covered agent subsets (see Observation 3)

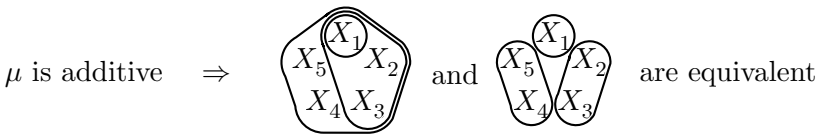


Figure 5: When additive, measurements of nested agent subsets can be equivalently defined as measurements of disjoint agent subsets (see Observation 4)

Observation 3. *For any* additive *generic measurement μ, the measurement of a partition refines the measurement of the covered subset:*

$$\forall A_1 \subset \Omega, \quad \forall A_2 \subset \Omega, \quad A_1 \cap A_2 = \emptyset \quad \Rightarrow \quad (\mu_{A_1}, \mu_{A_2}) \prec \mu_{A_1 \cup A_2}.$$

Proof. This directly follows the definitions of additive measurements and of the refinement relation (see Definitions 3 and 5). □

Observation 4. *For any* additive *generic measurement μ, the joint measurement of two nested agent subsets is equivalent to the measurement of the corresponding disjoint subsets:*

$$\forall A_1 \subset \Omega, \quad \forall A_2 \subset \Omega, \quad A_1 \subset A_2 \quad \Rightarrow \quad \begin{array}{ll} (\mu_{A_1}, \mu_{A_2}) & \prec \ (\mu_{A_1}, \mu_{A_2 \setminus A_1}) \\ (\mu_{A_1}, \mu_{A_2}) & \succ \ (\mu_{A_1}, \mu_{A_2 \setminus A_1}). \end{array}$$

Proof. Since $H(\mu_{A_2}(X) | \mu_{A_2 \setminus A_1}(X), \mu_{A_1}(X)) = 0$ (additivity), then $(\mu_{A_1}, \mu_{A_2 \setminus A_1})(X)$ fully determines $(\mu_{A_1}, \mu_{A_2})(X)$ and conversely, since $H(\mu_{A_2 \setminus A_1}(X) | \mu_{A_2}(X), \mu_{A_1}(X)) = 0$ (additivity), then $(\mu_{A_1}, \mu_{A_2})(X)$ fully determines $(\mu_{A_1}, \mu_{A_2 \setminus A_1})(X)$. □

Observation 5. *For any* additive *generic measurement μ, any feasible measurement* $(\mu_{A_1}, \ldots, \mu_{A_k})$ refines *the empty measurement $\mu_\emptyset$ and* is refined *by the microscopic measurement* $(\mu_{\{1\}}, \ldots, \mu_{\{N\}})$*. Hence, the empty measurement is the top element of the measurement poset (it has the lowest complexity and the lowest predictive capacity) and the microscopic measurement is the bottom element of the measurement poset and (it has highest complexity and the highest predictive capacity).*

Proof. First, since $\forall A \subset \Omega$ we have $H(\mu_\emptyset(X)|\mu_A(X)) = 0$ (additivity), then any feasible measurement $(\mu_{A_1}, \ldots, \mu_{A_k})(X)$ fully determines the empty measurement $\mu_\emptyset(X)$. Second, since $\forall A \subset \Omega$, we have $H(\mu_A(X)|(\mu_{\{i\}}(X))_{i\in A}) = 0$ (additivity), then any feasible measurement $(\mu_{A_1}, \ldots, \mu_{A_k})(X)$ is fully determined by the microscopic measurement $(\mu_{\{1\}}, \ldots, \mu_{\{N\}})(X)$. □

3.2 A General Result on Optimality of Nested Measurements

Under some hypotheses – presented in further details below – one can avoid evaluating all feasible measurements when solving the OPP. In this subsection, we indeed present some results allowing to only consider a subset of the solution space and thus considerably reducing the computation cost of the optimisation problem. Intuitively, (1) if the generic measurement μ is additive and deterministic, (2) if we know that some pre-measurement $\mu_\mathcal{A}$ contains all the information available at the microscopic level that is relevant to predict some post-measurement ψ, and (3) if the probability distribution $\Pr(X_A^t|\mu_\mathcal{A}(X_A^t))$ of the microscopic state given this pre-measurement is uniform, then, for any pre-measurement $\mu_\mathcal{B}$ such that $\mathcal{B}$ is "nested" in $\mathcal{A}$, one does not increase the predictive capacity of $\mu_\mathcal{B}$ by considering a partition of $\mathcal{B}$ instead of $\mathcal{B}$ itself. Hence, all partitions of collections of agent subsets that are "nested" in $\mathcal{A}$ can be removed from the solution space.

In Subsection 4.3, we apply this result to one of our case study: the prediction of the macroscopic aggregated-state of the Voter Model in the case of a complete interaction graph and a uniform initial distribution. Hence, thanks to this result, we are able to give a *complete characterisation* of the IB-diagram for this case study, that is a complete solution of the OPP for the parameter space $(t, \tau, \beta) \in \mathbb{N}\times\mathbb{N}\times\mathbb{R}^+$.

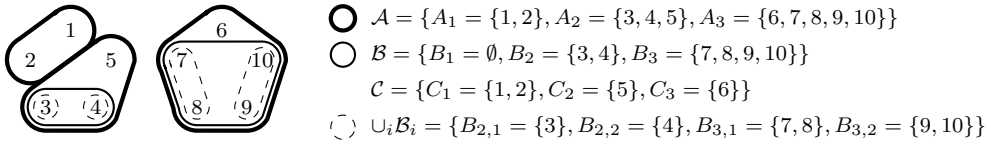


Figure 6: Example of setting for Theorems 2, 3 and 4 with ten agents and four collections of agent subsets $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ and $\cup_i \mathcal{B}_i$ as described in Subsection 3.2.2

3.2.1 Definitions and Notations

The following objects are illustrated in Fig. 6 by an example of feasible measurements satisfying these definitions.

- Given a generic measurement μ and a post-measurement ψ that one wants to predict;

- Given a collection of agent subsets $\mathcal{A} = \{A_1, \ldots, A_k\}$ that are pairwise disjoint[2]: $\forall i \neq j,\ A_i \cap A_j = \emptyset$;
- Given a collection of agent subsets $\mathcal{B} = \{B_1, \ldots, B_k\}$ that are each included in an agent subset in $\mathcal{A}$: $\forall i,\ B_i \subset A_i$.
- Given k collections of agent subsets $\mathcal{B}_i = \{B_{i,1}, \ldots, B_{i,l_i}\}$ that are each a partition of an agent subset in $\mathcal{B}$:

$$\forall i,\ B_{i,1} \cup \ldots \cup B_{i,l_i} = B_i \quad \text{and} \quad \forall i,\ \forall j_1 \neq j_2,\ B_{i,j_1} \cap B_{i,j_2} = \emptyset.$$

In the following, we also mark:

- $\forall i,\ C_i = A_i \backslash B_i$ the subset of agents covered by A_i but not by B_i, and $\mathcal{C} = \{C_1, \ldots, \mathcal{C}_k\}$ the collection of these subsets;
- $A = A_1 \cup \ldots \cup A_k$ the set of agents covered by $\mathcal{A}$;
- $B = B_1 \cup \ldots \cup B_k$ the set of agents covered by $\mathcal{B}$;
- $C = A \backslash B$ the set of agents covered by $\mathcal{A}$ but not by $\mathcal{B}$;
- $X_A \in S^{|A|}$, $X_B \in S^{|B|}$, and $X_C \in S^{|C|}$ the microscopic state of the corresponding agents at time t. For example, $X_A = (X_i^t)_{i \in A}$. Hence, we have $X_A = X_B \cup X_C$.

3.2.2 Assumptions and General Reasoning

(A1) If μ is *additive* and *deterministic*;

(A2) If the feasible pre-measurement $\mu_{\mathcal{A}} = (\mu_{A_1}, \ldots, \mu_{A_k})$ resulting from collection $\mathcal{A}$ contains all the information that is available at the microscopic level to predict ψ:

$$\forall t \in \mathbb{N},\ \forall \tau \in \mathbb{N}, \quad I(X^t; \psi(X^{t+\tau}) | \mu_{\mathcal{A}}(X^t)) = 0;$$

(A3) If the microscopic Markov chain has the *uniform micro-state property* with respect to $\mu_{\mathcal{A}}$, that is that, at any time, the probability distribution of X_A^t given $\mu_{\mathcal{A}}(X_A^t)$ is uniform: $\forall t \in \mathbb{N},\ \forall x_A \in S^{|A|},\ \forall x'_A \in S^{|A|},\ \forall s_\mu \in (\mathcal{S}_\mu)^k$,

$$\Pr(X_A^t = x_A | \mu_{\mathcal{A}}(X_A^t) = s_\mu) \ = \ \Pr(X_A^t = x'_A | \mu_{\mathcal{A}}(X_A^t) = s_\mu),$$

and, because μ is deterministic: $\forall t \in \mathbb{N},\ \forall x_A \in S^{|A|},\ \forall x'_A \in S^{|A|}$,

$$\mu_{\mathcal{A}}(x_A) = \mu_{\mathcal{A}}(x'_A) \quad \Rightarrow \quad \Pr(X_A^t = x_A) = \Pr(X_A^t = x'_A);$$

(Result) Then, the feasible pre-measurement $\mu_{\mathcal{B}} = (\mu_{B_1}, \ldots, \mu_{B_k})$ resulting from collection $\mathcal{B}$ is always more efficient than the feasible pre-measurement $\mu_{\cup_i \mathcal{B}_i} = (\mu_{B_{1,1}}, \ldots, \mu_{B_{1,l_1}}, \ldots, \mu_{B_{k,1}}, \ldots, \mu_{B_{k,l_k}})$ resulting from collection $\cup_i \mathcal{B}_i$.

[2]Note that, when μ is additive, this hypothesis can also be applied to collections of nested agent subsets (with possible $A_i \subset A_j$) by defining the corresponding collection of disjoint agent subsets (see Observation 4).

3.2.3 Theorems

Given that the previous assumptions hold, then we have the three following theorems.

Theorem 2. *At any time, $\mu_{\mathcal{B}}$ has a lower complexity than $\mu_{\cup_i \mathcal{B}_i}$:*

$$\forall t \in \mathbb{N}, \quad I(X^t; \mu_{\mathcal{B}}(X^t)) \quad \leq \quad I(X^t; \mu_{\cup_i \mathcal{B}_i}(X^t)).$$

Theorem 3. *At any time and for any horizon, $\mu_{\mathcal{B}}$ and $\mu_{\cup_i \mathcal{B}_i}$ have the same predictive capacity:*

$$\forall t \in \mathbb{N}, \ \forall \tau \in \mathbb{N}, \quad I(\mu_{\mathcal{B}}(X^t); \psi(X^{t+\tau})) \quad = \quad I(\mu_{\cup_i \mathcal{B}_i}(X^t); \psi(X^{t+\tau})).$$

Theorem 4. *At any time and for any horizon, the β-border between the optimality region of $\mu_{\mathcal{B}}$ and the optimality region of $\mu_{\cup_i \mathcal{B}_i}$ is either infinite or defined everywhere. Hence, for any value of the trade-off parameter, $\mu_{\cup_i \mathcal{B}_i}$ is never more efficient than $\mu_{\mathcal{B}}$:*

$\forall t \in \mathbb{N}, \ \forall \tau \in \mathbb{N}, \ \forall \beta \in \mathbb{R}^+,$

$$IB_\beta(X^t; \mu_{\mathcal{B}}(X^t); \psi(X^{t+\tau})) \quad \leq \quad IB_\beta(X^t; \mu_{\cup_i \mathcal{B}_i}(X^t); \psi(X^{t+\tau})).$$

The proof of Theorem 3 requires the following two lemmas.

Lemma 1. *$\forall x_B \in S^{|B|}$ and $\forall x'_B \in S^{|B|}$ such that $\mu_{\mathcal{B}}(x_B) = \mu_{\mathcal{B}}(x'_B) = b \in (\mathcal{S}_\mu)^k$ and $\forall x_C \in S^{|C|}$ such that $\mu_{\mathcal{C}}(x_C) = c \in (\mathcal{S}_\mu)^k$, we have:*

(1) $\Pr(X_B^t = x_B, X_C^t = x_C) \ = \ \Pr(X_B^t = x'_B, X_C^t = x_C)$;

(2) $\Pr(X_B^t = x_B) \ = \ \Pr(X_B^t = x'_B)$;

(3) $\Pr(\mu_{\mathcal{C}}(X_C^t) = c | X_B^t = x_B) \ = \ \Pr(\mu_{\mathcal{C}}(X_C^t) = c | \mu_{\mathcal{B}}(X_B^t) = b)$;

(4) $H(\mu_{\mathcal{C}}(X_C^t) | X_B^t) \ = \ H(\mu_{\mathcal{C}}(X_C^t) | \mu_{\mathcal{B}}(X_B^t))$;

(5) $I(\mu_{\cup_i \mathcal{B}_i}(X^t); \mu_{\mathcal{A}}(X^t) | \mu_{\mathcal{B}}(X^t)) \ = \ 0$;

Lemma 2. *$\forall x_B \in S^{|B|}$ and $\forall x'_B \in S^{|B|}$ such that $\mu_{\mathcal{B}}(x_B) = \mu_{\mathcal{B}}(x'_B) = b \in (\mathcal{S}_\mu)^k$, $\forall x_C \in S^{|C|}$ such that $\mu_{\mathcal{C}}(x_C) = c \in (\mathcal{S}_\mu)^k$, and $\forall s \in \mathcal{S}_\psi$, we have:*

(1) $\Pr(X_B^t = x_B, X_C^t = x_C, \psi(X^{t+\tau}) = s)$
$= \ \Pr(X_B^t = x'_B, X_C^t = x_C, \psi(X^{t+\tau}) = s)$;

(2) $\Pr(X_B^t = x_B, \psi(X^{t+\tau}) = s) \ = \ \Pr(X_B^t = x'_B, \psi(X^{t+\tau}) = s)$;

(3) $\Pr(\mu_{\mathcal{C}}(X_C^t) = c | X_B^t = x_B, \psi(X^{t+\tau}) = s)$
$= \ \Pr(\mu_{\mathcal{C}}(X_C^t) = c | \mu_{\mathcal{B}}(X_B^t) = b, \psi(X^{t+\tau}) = s)$;

(4) $H(\mu_{\mathcal{C}}(X_C^t) | X_B^t, \psi(X^{t+\tau})) \ = \ H(\mu_{\mathcal{C}}(X_C^t) | \mu_{\mathcal{B}}(X_B^t), \psi(X^{t+\tau}))$;

(5) $I(\mu_{\cup_i \mathcal{B}_i}(X^t); \mu_{\mathcal{A}}(X^t) | \mu_{\mathcal{B}}(X^t), \psi(X^{t+\tau})) \ = \ 0$.

3.2.4 Proofs

For simplicity, in the following proofs, we omit the time index t when non-ambiguous. For example, we mark X_A instead of X_A^t.

Proof of Lemma 1.
(1) $\forall i \in \{1, \ldots, k\}$, since μ is additive and deterministic, and since $B_i \cup C_i = A_i$ and $B_i \cap C_i = \emptyset$, we have $H(\mu_{A_i}(X_B, X_C)|\mu_{B_i}(X_B), X_C) = 0$ (see Observation 1). Hence, $H(\mu_{\mathcal{A}}(X_B, X_C)|\mu_{\mathcal{B}}(X_B), X_C) = 0$. Therefore, if $\mu_{\mathcal{B}}(X_B)$ and X_C are fixed, then $\mu_{\mathcal{A}}(X_B, X_C)$ is also fixed. So, in our case, since $\mu_{\mathcal{B}}(x_B) = \mu_{\mathcal{B}}(x'_B)$, we have $\mu_{\mathcal{A}}(x_B, x_C) = \mu_{\mathcal{A}}(x'_B, x_C)$. Hence, because of the uniform micro-state property (A3), we have

$$\Pr(X_B = x_B, X_C = x_C) \quad = \quad \Pr(X_B = x'_B, X_C = x_C).$$

$$\begin{aligned}(2)\quad \Pr(X_B = x_B) &= \sum_{x_C \in S^{|C|}} \Pr(X_B = x_B, X_C = x_C) \\ &= \sum_{x_C \in S^{|C|}} \Pr(X_B = x'_B, X_C = x_C) \qquad \text{(see (1))} \\ &= \Pr(X_B = x'_B).\end{aligned}$$

$$\begin{aligned}(3)\quad \Pr(\mu_{\mathcal{C}}(X_C) = c|X_B = x_B) &= \sum_{\substack{x_C \in S^{|C|} \\ \mu_{\mathcal{C}}(x_C) = c}} \frac{\Pr(X_B = x_B, X_C = x_C)}{\Pr(X_B = x_B)} \\ &= \sum_{\substack{x_C \in S^{|C|} \\ \mu_{\mathcal{C}}(x_C) = c}} \frac{\Pr(X_B = x'_B, X_C = x_C)}{\Pr(X_B = x'_B)} \quad \text{(see (1) and (2))} \\ &= \Pr(\mu_{\mathcal{C}}(X_C) = c|X_B = x'_B).\end{aligned}$$

Therefore, conditioning $\mu_{\mathcal{C}}(X_C)$ on X_B is equivalent to conditioning $\mu_{\mathcal{C}}(X_C)$ on $\mu_{\mathcal{B}}(X_B)$:

$$\Pr(\mu_{\mathcal{C}}(X_C) = c|X_B = x_B) \quad = \quad \Pr(\mu_{\mathcal{C}}(X_C) = c|\mu_{\mathcal{B}}(X_B) = b).$$

(4) First, following (3), we have

$$H(\mu_{\mathcal{C}}(X_C)|X_B = x_B) \quad = \quad H(\mu_{\mathcal{C}}(X_C)|\mu_{\mathcal{B}}(X_B) = b).$$

Second,

$$H(\mu_{\mathcal{C}}(X_C)|X_B) = \sum_{x_B \in S^{|B|}} \Pr(X_B = x_B)\ H(\mu_{\mathcal{C}}(X_C)|X_B = x_B)$$

$$
\begin{aligned}
&= \sum_{b \in (\mathcal{S}_\mu)^k} \left(\sum_{\substack{x_B \in S^{|B|} \\ \mu_{\mathcal{B}}(x_B) = b}} \Pr(X_B = x_B) \right) H(\mu_{\mathcal{C}}(X_C) | \mu_{\mathcal{B}}(X_B) = b) \\
&= \sum_{b \in (\mathcal{S}_\mu)^k} \Pr(\mu_{\mathcal{B}}(X_B) = b)\ H(\mu_{\mathcal{C}}(X_C) | \mu_{\mathcal{B}}(X_B) = b) \\
&= H(\mu_{\mathcal{C}}(X_C) | \mu_{\mathcal{B}}(X_B))
\end{aligned}
$$

(5) $\forall i \in \{1, \ldots, k\}$, since μ is additive and deterministic, $B_i \cup C_i = A_i$ and $B_i \cap C_i = \emptyset$, we have $H(\mu_{A_i}(X_B, X_C) | X_B, \mu_{C_i}(X_C)) = 0$ and $H(\mu_{C_i}(X_C) | \mu_{A_i}(X_B, X_C), X_B) = 0$ (see Observation 1). Hence, $H(\mu_{\mathcal{A}}(X) | X_B, \mu_{\mathcal{C}}(X)) = 0$ and $H(\mu_{\mathcal{C}}(X) | \mu_{\mathcal{A}}(X), X_B) = 0$.

Then, since

$$
\begin{aligned}
I(\mu_{\mathcal{A}}(X); \mu_{\mathcal{C}}(X) | X_B) &= H(\mu_{\mathcal{A}}(X) | X_B) \ - \ H(\mu_{\mathcal{A}}(X) | X_B, \mu_{\mathcal{C}}(X)) \\
&= H(\mu_{\mathcal{C}}(X) | X_B) \ - \ H(\mu_{\mathcal{C}}(X) | \mu_{\mathcal{A}}(X), X_B),
\end{aligned}
$$

we have $H(\mu_{\mathcal{C}}(X) | X_B) = H(\mu_{\mathcal{A}}(X) | X_B)$, and the same reasoning also gives $H(\mu_{\mathcal{C}}(X) | \mu_{\mathcal{B}}(X)) = H(\mu_{\mathcal{A}}(X) | \mu_{\mathcal{B}}(X))$.

Hence, following (4), we have $H(\mu_A(X) | X_B) = H(\mu_A(X) | \mu_{\mathcal{B}}(X))$, and then

$$
I(X_B; \mu_{\mathcal{A}}(X) | \mu_{\mathcal{B}}(X)) \ = \ H(\mu_{\mathcal{A}}(X) | \mu_{\mathcal{B}}(X)) - H(\mu_{\mathcal{A}}(X) | X_B) \ = \ 0.
$$

Hence, $X_B \to \mu_{\mathcal{B}}(X) \to \mu_{\mathcal{A}}(X)$ is a Markov chain and, since $\mu_{\cup_i \mathcal{B}_i}(X) \to X_B \to \mu_{\mathcal{B}}(X)$ is a also Markov chain, then we finally have

$$
I(\mu_{\cup_i \mathcal{B}_i}(X^t); \mu_{\mathcal{A}}(X^t) | \mu_{\mathcal{B}}(X^t)) = 0.
$$

□

Proof of Lemma 2. The proof of Lemma 2 takes the exact same form as the one of Lemma 1 by also considering that, because $I(X_B, X_C; \psi(X^{t+\tau}) | \mu_{\mathcal{A}}(X^t)) = 0$, we have, $\forall s \in \mathcal{S}_\psi$,

$$
\begin{aligned}
\Pr(\psi(X^{t+\tau}) = s | X_B = x_B, X_C = x_C) &= \Pr(\psi(X^{t+\tau}) = s | \mu_{\mathcal{A}}(x_B, x_C)) \\
&= \Pr(\psi(X^{t+\tau}) = s | \mu_{\mathcal{A}}(x'_B, x_C)) \\
&= \Pr(\psi(X^{t+\tau}) = s | X_B = x'_B, X_C = x_C).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\Pr(X_B = x_B, X_C = x_C, \psi(X^{t+\tau}) = s) \\
&= \Pr(X_B = x_B, X_C = x_C)\ \Pr(\psi(X^{t+\tau}) = s | X_B = x_B, X_C = x_C) \\
&= \Pr(X_B = x'_B, X_C = x_C)\ \Pr(\psi(X^{t+\tau}) = s | X_B = x'_B, X_C = x_C) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(also see (1) of Lemma 1)}
\end{aligned}
$$

$= \Pr(X_B = x'_B, X_C = x_C, \psi(X^{t+\tau}) = s)$.

Then, one can applies the same steps as in (2), (3), (4), and (5) above to show

$$I(\mu_{\cup_i \mathcal{B}_i}(X^t); \mu_{\mathcal{A}}(X^t) | \mu_{\mathcal{B}}(X^t), \psi(X^{t+\tau}) = 0.$$

□

Proof of Theorem 2. Since μ is additive, and because $\cup_i \mathcal{B}_i$ is a partition of $\mathcal{B}$, we have $\mu_{\cup_i \mathcal{B}_i} \prec \mu_{\mathcal{B}}$ (see Observation 3). Then, $\forall t \in \mathbb{N}$, $I(X^t; \mu_{\mathcal{B}}(X^t)) \leq I(X^t; \mu_{\cup_i \mathcal{B}_i}(X^t))$ (see Theorem 1). □

Proof of Theorem 3. In this proof, we simply notations as follows: we mark $\mathcal{A}$ instead of $\mu_{\mathcal{A}}(X^t)$; $\mathcal{B}$ instead of $\mu_{\mathcal{B}}(X^t)$; $\cup_i \mathcal{B}_i$ instead of $\mu_{\cup_i \mathcal{B}_i}(X^t)$; and ψ instead of $\psi(X^{t+\tau})$. Then, we use the following conditional independences:

(I1) $(\mathcal{A}, \mathcal{B}, \cup_i \mathcal{B}_i) \to X \to \psi$ because the future states, and so the future post-measurements, fully depend on the current microscopic state;

(I2) $X \to \mathcal{A} \to \psi$ according to (A2);

(I3) $(\mathcal{B}, \cup_i \mathcal{B}_i) \to \mathcal{A} \to \psi$ directly follows from (I1) and (I2);

(I4) $\cup_i \mathcal{B}_i \to \mathcal{B} \to \mathcal{A}$ according to result (5) of Lemma 1.

(I5) $\cup_i \mathcal{B}_i \to (\mathcal{B}, \psi) \to \mathcal{A}$ according to result (5) of Lemma 2.

The inequality holds if $\cup_i \mathcal{B}_i \to \mathcal{B} \to \psi$ is a Markov chain. Indeed, in this case, the Data Processing Inequality gives: $I(\mathcal{B}; \psi) \geq I(\cup_i \mathcal{B}_i; \psi)$. Hence, we want to show that $I(\psi; \cup_i \mathcal{B}_i | \mathcal{B}) = 0$. We have:

$$\begin{aligned}
I(\psi; \cup_i \mathcal{B}_i | \mathcal{B}) &= I(\psi; \cup_i \mathcal{B}_i, X, \mathcal{A} | \mathcal{B}) - I(\psi; X, \mathcal{A} | \mathcal{B}, \cup_i \mathcal{B}_i) && \text{(chain rule)} \\
&= I(\psi; \mathcal{A} | \mathcal{B}) - I(\psi; \mathcal{A} | \mathcal{B}, \cup_i \mathcal{B}_i) && \text{(from I2 and I3)} \\
&= I(\psi, \cup_i \mathcal{B}_i; \mathcal{A} | \mathcal{B}) - I(\cup_i \mathcal{B}_i; \mathcal{A} | \mathcal{B}, \psi) \\
&\quad + I(\cup_i \mathcal{B}_i; \mathcal{A} | \mathcal{B}) - I(\psi, \cup_i \mathcal{B}_i; \mathcal{A} | \mathcal{B}) && \text{(two chain rules)} \\
&= 0 && \text{(from I4)}
\end{aligned}$$

Moreover, since $\mathcal{B}$ is a partition of B, and since μ is additive, we have $\mu_{\cup_i \mathcal{B}_i} \prec \mu_{\mathcal{B}}$ (see Observation 3). Then, $\forall t \in \mathbb{N}$, $\forall \tau \in \mathbb{N}$, $I(\mu_{\mathcal{B}}(X^t); \psi(X^{t+\tau})) \leq I(\mu_{\cup_i \mathcal{B}_i}(X^t); \psi(X^{t+\tau}))$ (see Theorem 1). Hence, we have an equality. □

Proof of Theorem 4. Directly follows Theorems 2 and 3 (also see Table 1). □

3.2.5 Generalisation and Conjectures

In Subsections 4.3, 5.1, and 5.2, we will apply these theorems to practical cases in order to efficiently solve the OPP by enumerating and evaluating only the pre-measurements that are not declared "non-optimal" by Theorem 4. However, in these practical cases, it seems that the three theorems actually apply to a broader class of pre-measurements, thus allowing to reduce even more the list of possible solutions. Hence, we present in this subsection a generalisation of the three theorems to this broader class. As we do not have a proof yet for this generalisation, we simply formulate three conjectures while recording no empirical violation in the practical examples we considered.

Thus, we conjecture that Theorems 2, 3 and 4 also applies when, $\forall i \in \{1, \ldots, k\}$, $\mathcal{B}_i$ is any collection of agent subsets covering B_i, and not necessarily a *partition* of B_i (we might have $B_{i,j_1} \cap B_{i,j_2} \neq \emptyset$). In this way, one would be able to reduce the solution space further.

In other words, if we would replace the definition of $\mathcal{B}_i$ in Subsection 3.2.1 by:

- Given k collections of agent subsets $\mathcal{B}_i = \{B_{i,1}, \ldots, B_{i,l_i}\}$ that each covers an agent subset in $\mathcal{B}$: $\quad \forall i,\ B_{i,1} \cup \ldots \cup B_{i,l_i} = B_i$,

then we would have analogous results as the ones of Theorems 2, 3 and 4.

Conjecture 1. *At any time, $\mu_{\mathcal{B}}$ has a lower complexity than $\mu_{\cup_i \mathcal{B}_i}$:*

$$\forall t \in \mathbb{N}, \quad I(X^t; \mu_{\mathcal{B}}(X^t)) \quad \leq \quad I(X^t; \mu_{\cup_i \mathcal{B}_i}(X^t)).$$

Among the three conjectures of this subsection, this first one seems to be the more difficult to prove. Indeed, it requires (1) looking in details at the way the microscopic state space is aggregated when one measures non-disjoint agent subsets, (2) comparing the result to the case where one instead measures disjoint agent subsets, (3) showing that the uniform micro-state property implies that the complexity of the former case is always larger than the complexity of the latter case.

Conjecture 2. *At any time and for any horizon, $\mu_{\mathcal{B}}$ has a higher predictive capacity than $\mu_{\cup_i \mathcal{B}_i}$:*

$$\forall t \in \mathbb{N},\ \forall \tau \in \mathbb{N}, \quad I(\mu_{\mathcal{B}}(X^t); \psi(X^{t+\tau})) \quad \geq \quad I(\mu_{\cup_i \mathcal{B}_i}(X^t); \psi(X^{t+\tau})).$$

Indeed, Lemmas 1 and 2 would also apply in this more general case, and the above inequality is actually proved in the first part of the proof of Theorem 3.

Conjecture 3. *At any time and for any horizon, the β-border between the optimality region of $\mu_{\mathcal{B}}$ and the optimality region of $\mu_{\cup_i \mathcal{B}_i}$ is either null, infinite, defined nowhere or defined everywhere (*i.e.*, it is never strictly positive). Hence, for any value of the trade-off parameter, $\mu_{\cup_i \mathcal{B}_i}$ is never more efficient than $\mu_{\mathcal{B}}$:*

$\forall t \in \mathbb{N},\ \forall \tau \in \mathbb{N},\ \forall \beta \in \mathbb{R}^+$,

$$IB_\beta(X^t; \mu_{\mathcal{B}}(X^t); \psi(X^{t+\tau})) \quad \leq \quad IB_\beta(X^t; \mu_{\cup_i \mathcal{B}_i}(X^t); \psi(X^{t+\tau})).$$

This result would directly follow from Conjectures 1 and 2 (also see Table 1).

4 Application to the Voter Model

We now apply our framework on a quite simple – yet illustrative – ABM, namely the Voter Model (VM). Originally developed in the context of population genetics [19, 23] and species competition [8], the VM has today become a canonical example for interacting particle systems [21] and a standard model in the area of opinion and social dynamics [7]. Though it is probably too simple for drawing direct conclusions related to real applications, the VM is completely suited for our purposes since it entails the possibility to introduce heterogeneity and multilevel organisation in ABMs while, at the same time, it is simple enough to work with an explicit representation of the microscopic transition matrix and to directly compute the IB measures defined in Section 2.

4.1 Model Presentation

In this paper, use the definitions of the VM and the corresponding microscopic transition rates as defined in previous work [4]. Each agent can be in two possible states $S = \{0, 1\}$ and can synchronise its state with other agents according to a directed interaction graph. The dynamics then corresponds to a sequential update: Each step of the system's dynamics hence consists in the random selection of a directed edge (i, j) and the update of the state of agent j according to the state of agent i, all other agents staying in the same state (see Fig. 7 for an example of such dynamics):

$$\text{edge } (i,j) \text{ selected at time } t \quad \Rightarrow \quad X_j^{t+1} = X_i^t \quad \text{and} \quad \forall k \neq j,\ X_k^{t+1} = X_k^t.$$

When the interaction graph is connected, the VM has two trivial attractors: The agents all end up in state 0 or in state 1. Hence, the system converges to an homogeneous state, and the OPP hence depends on the time t at which the pre-measurement is performed and on the horizon τ defining the time $t + \tau$ at which the post-measurement that one wants to predict is performed. In the following, we assume that the system starts in a fully random state:

$$\forall x \in \Sigma,\ \Pr(X^0 = x) = 2^{-N} \quad \Rightarrow \quad \forall i \in \Omega,\ \Pr(X_i^0 = 0) \ = \ \Pr(X_i^0 = 1) \ = \ 1/2.$$

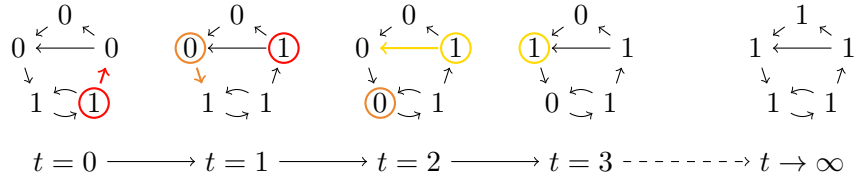


Figure 7: Example of dynamics of a five-agent Voter Model

The Aggregated-state Generic Measurement. In our experiments, we consider the following canonical generic measurement (and the corresponding feasible measurements) for the VM.

Definition 6 (Aggregated-state Measurement)**.** *The* aggregated-state measurement η *is an additive generic measurement from* $\{0,1\}^N$ *to* $(\mathbb{N},+)$ *that simply performs a summation of the agent states:*

$$\forall A \subset \Omega, \quad \eta_A(X) = \sum_{i \in A} X_i.$$

The aggregated-state measurement thus indicates the number of agents in state 1 within the considered agent subset. Note that it is deterministic and additive with respect to the addition $+$ in $\mathbb{N}$:

$$\forall A_1 \subset \Omega,\ \forall A_2 \subset \Omega, \quad A_1 \cap A_2 = \emptyset \quad \Rightarrow \quad \eta_{A_1 \cup A_2} = \eta_{A_1} + \eta_{A_2}.$$

4.2 Numerical Approximation of the IB-variational

The experiments presented in the following sections have required the development of a dedicated *numerical approximation* program, implemented in `C++` and freely available on a GitHub repository [20]. Contrary to *simulation-based estimation* methods, the IB-measures are here approximated by computing the "exact" probability distributions of the random variables of interest (with basic matrix calculus). More precisely, the program takes inputs from a directed interaction graph (*i.e.*, a set of N nodes and a set of edges with weights), an initial distribution $\Pr(X^0)$ over the microscopic state space $\{0,1\}^N$, a time t, an horizon τ, a pre-measurement ϕ and a post-measurement ψ (*i.e.*, two partitions of the microscopic state space that are actually derived from a generic measurement μ and two collections of agent subsets). From this input, the program computes the transition matrix $T(X^{t+1}|X^t)$ of the corresponding Markov chain on the microscopic state space, and the following probability distributions: $\Pr(X^t)$, $\Pr(\phi(X^t))$, $\Pr(X^{t+\tau})$, $\Pr(\psi(X^{t+\tau}))$, $\Pr(\phi(X^t),\psi(X^{t+\tau}))$. Finally, the program uses these probability distributions to compute the two IB-measures, that is the pre-measurement complexity $I(X^t;\phi(X^t))$ and its predictive capacity $I(\phi(X^t);\psi(X^{t+\tau}))$.

Hence, except for Subsection 4.3, our results do not rely on any *analytical* expression of the IB-measures, but only on *numerical* approximations of the transition matrix. However, such results are numerically *exact*, as opposed to numerical *estimation*, since the probability distributions of random variables are not estimated by simulation, but fully computed from the initial transition matrix. By taking into account the accuracy of the double-precision floating-point format in `C++`, we can typically guaranty 16 significant digits[3].

[3]This guarantee holds only if the inputs are themselves within this accuracy range. Moreover, because of the possible summation of rounding errors, this accuracy becomes lower for large systems, large times and large horizons. In practice, we easily guarantee a 10-digits accuracy of all experiments in this paper.

This program is generic in the sense that it can be executed on any interaction graph. However, because the size of the state space exponentially depends on the number of agents, we are strongly limited in the system size. In practice, we manage to scale up to 12 agents (the transition matrix then contains 16.7 millions cells). To overcome this limitation, we also implemented a "compact model" of the two-community case (see Subsection 5.2). In this model, the state space is lumped according to the multilevel measurement **(AGENT+MESO1+MACRO)**, as defined in Subsection 5.2 and as represented in the bottom-right corner of Fig. 12. This measurement indeed contains all the information that is actually required to compute the IB-variational for the feasible measurements that we consider in the following (see the lumpable partition of the two-community VM in [4]). In this compact model, the size of the state space linearly depends on the size of each community ($2\ (N_1 - 1)\ N_2$ possible states, where N_1 and N_2 are the respective size of the two communities) and it can easily be scaled up to 2 times 20 agents.

4.3 Predicting the Macroscopic Measurement in the Complete Graph

The most simple setting of the VM corresponds to a complete and uniform interaction graph: all agents are connected one to another and all edges are equally likely to be selected at each simulation step for synchronisation. The resulting global dynamics, $\forall t \in \mathbb{N}$, is thus the following: $\forall x = (x_1, \ldots, x_N) \in S^N, \forall x' = (x'_1, \ldots, x'_N) \in S^N$,

$T\left(X^{t+1} = x' | X^t = x\right) =$

$$\begin{cases} \frac{\overline{x}'-1}{N\ (N-1)} & \text{if } \exists i,\ x_i = 0,\ x'_i = 1 \text{ and } \forall j \neq i,\ x_j = x'_j, \quad \text{(case 1)} \\ \frac{N-(\overline{x}'+1)}{N\ (N-1)} & \text{if } \exists i,\ x_i = 1,\ x'_i = 0 \text{ and } \forall j \neq i,\ x_j = x'_j, \quad \text{(case 2)} \\ \frac{(N-\overline{x}')^2+(\overline{x}')^2-N}{N\ (N-1)} & \text{if } \forall j,\ x_j = x'_j, \quad \text{(case 3)} \\ 0 & \text{else,} \end{cases}$$

with $\overline{x}' = \sum_{i \in \Omega} x'_i$.

Applying the General Result. Our first experiment deals with the optimal prediction of the macroscopic aggregated-state measurement η_Ω in the complete graph. In this case, we can apply the general result of Subsection 3.2 to provide an optimal solution of the OPP for the aggregated-state measurement.

(A1) As stated above, η is both additive and deterministic.

(A2) It has been shown that the macroscopic aggregated-state measurement is *lumpable* in the case of a complete and uniform interaction graph [2], meaning in particular that it contains all the information available at the microscopic level regarding its own dynamics:

$$I(X^t; \eta_\Omega(X^{t+\tau}) | \eta_\Omega(X^t)) = 0.$$

Hence, η_Ω is fully informative and satisfy assumption (A2) of Subsection 3.2.

(A3) Moreover, the microscopic Markov chain have the uniform micro-state property with respect to the aggregated-state measurement. Indeed, this is true at time $t = 0$ since the system's state is uniformly distributed, and we can easily show that, if it is true at time $t \in \mathbb{N}$, then it is also true at time $t + 1$. $\forall x' \in S^N$,

$$\begin{aligned}
\Pr(X^{t+1} = x') \;&=\; \sum_{x \in S^N} T(X^{t+1} = x' | X^t = x)\ \Pr(X^t = x) \\
&= \left(\frac{\overline{x}' - 1}{N\,(N-1)}\right) \sum_{\text{case 1}} \Pr\left(X^t = x\right) \\
&\quad + \left(\frac{N - (\overline{x}' + 1)}{N\,(N-1)}\right) \sum_{\text{case 2}} \Pr\left(X^t = x\right) \\
&\quad + \left(\frac{(N - \overline{x}')^2 + (\overline{x}')^2 - N}{N\,(N-1)}\right) \sum_{\text{case 3}} \Pr\left(X^t = x\right) \qquad \text{(see formula above)} \\
&= \left(\frac{\overline{x}' - 1}{N\,(N-1)}\right) \overline{x}'\ \Pr\left(\eta_\Omega(X^t) = \overline{x}' - 1\right) \\
&\quad + \left(\frac{N - (\overline{x}' + 1)}{N\,(N-1)}\right) (N - \overline{x}')\ \Pr\left(\eta_\Omega(X^t) = \overline{x}' + 1\right) \\
&\quad + \left(\frac{(N - \overline{x}')^2 + (\overline{x}')^2 - N}{N\,(N-1)}\right) \Pr\left(\eta_\Omega(X^t) = \overline{x}'\right) \\
&\qquad\qquad \text{(because of the uniform micro-state property at time } t\text{)}
\end{aligned}$$

Since $\Pr(X^{t+1} = x')$ only depends on $\overline{x}'$ and not on x', we hence have

$$\eta_\Omega(x'_1) = \eta_\Omega(x'_2) \quad \Rightarrow \quad \overline{x}'_1 = \overline{x}'_2 \quad \Rightarrow \quad \Pr(X^{t+1} = x'_1) = \Pr(X^{t+1} = x'_2),$$

and the uniform micro-state property holds at time $t + 1$ (A3).

(Result) Consequently, in this setting, any measurement $(\eta_{B_1}, \ldots, \eta_{B_k})$ of a collection of disjoint agent subsets is never more efficient than the measurement $\eta_{B_1 \cup \ldots \cup B_k}$ of their union (see Theorem 4). Moreover, by using the conjecture expressed in Subsection 3.2.5, we also assume that this holds when agent subsets $B_1, \ldots, B_k$ are not disjoint.

Hence, one can consider only the following measurements when trying to solve the OPP on the complete graph (see Fig. 2 for a graphical representation of these feasible measurements):

EMPTY	The empty measurement $\eta_\emptyset$;
AGENT	An agent measurement $\eta_{\{i\}}$;

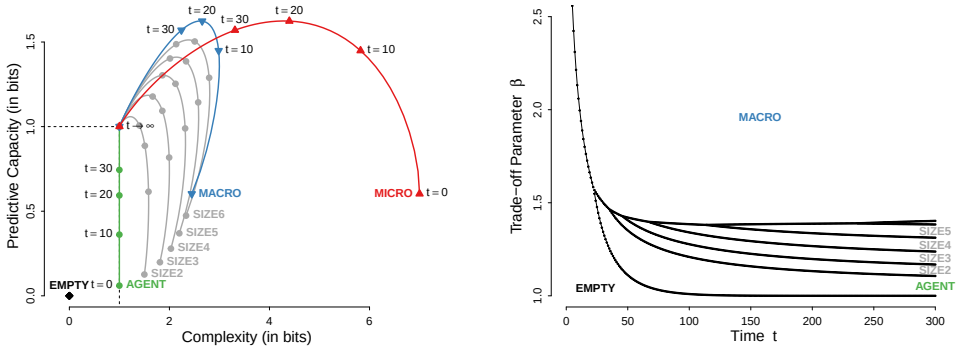


Figure 8: Predicting the macroscopic measurement in the complete graph (size $N = 7$, fixed horizon $\tau = 3$ and variable time $t \in \mathbb{N}$)

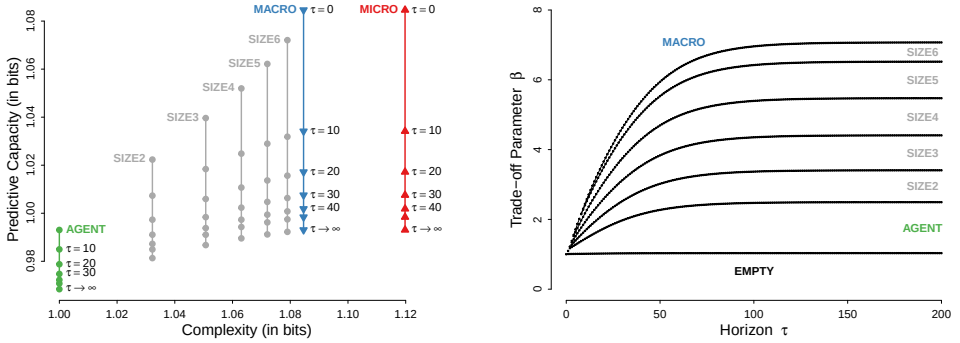


Figure 9: Predicting the macroscopic measurement in the complete graph (size $N = 7$, fixed time $t = 100$ and variable horizon $\tau \in \mathbb{N}$)

SIZE2	The measurement of a subset of two agents $\eta_{\{i,j\}}$;
SIZE3	The measurement of a subset of three agents $\eta_{\{i,j,k\}}$;
...	
MACRO	The macroscopic measurement $\eta_\Omega = \eta_{\{1,\dots,N\}}$.

Note that, because of the uniform micro-state property (A3), for any time t, the distribution probability $\Pr(X_i^t)$ of the state of agent i is the same as the distribution probability of the state of any other agent. Hence, the complexity and the predictive capacity of the agent measurement $\eta_{\{i\}}$ is the same for any other agent. This also holds for the complexity and the predictive capacity of any measurement η_B with $B \subset \Omega$ (see **SIZE** above), which only depend on the number of agent in B.

Fig. 8 and 9 present four graphs to illustrate our results on specific values of the three parameters. To be readable, all figures correspond to the complete graph of size $N = 7$. Fig. 8 corresponds to the prediction problem for a fixed horizon $\tau = 3$ and Fig. 9 for a fixed time $t = 100$. The two plots on the left give the complexity and the predictive capacity of feasible pre-measurements depending on the current time (Fig. 8) or depending on the horizon (Fig. 9). The two plots on the right represent bidimensional cuts of the tridimensional IB-diagram, that is the optimality regions of feasible pre-measurements for a fixed horizon ($\tau = 3$ in Fig. 8) and for a fixed time ($t = 100$ in Fig. 9).

Macro is Always Better than Micro. Since $\{\{1\}, \ldots, \{N\}\}$ is a (canonical) partition of Ω, Theorems 2, 3 and 4 can be applied to the microscopic and macroscopic measurements $(\eta_{\{1\}}, \ldots, \eta_{\{N\}})$ and η_Ω.

- By applying Theorem 2, we can see in the first plots of Fig. 8 and 9 that, at any time and for any horizon, the macroscopic measurement has a lower complexity than the microscopic measurement.
- By applying Theorem 3, we can see in the same plots that, at any time and for any horizon, the predictive capacities of both measurements are however the same.
- As a result, and by applying Theorem 4, we can deduce that the microscopic measurement will never be more efficient than the macroscopic measurement. In fact, the microscopic measurement becomes optimal only for an infinite value of the trade-off parameter, that is when the bottleneck variational is actually dominated by the prediction term and the complexity term hence becomes negligible. This is represented in the second plots of Fig. 8 and 9 by the fact that the microscopic measurement does not appear as optimal for finite values of β.
- The same reasoning can actually be applied to any couple of measurements $(\eta_{B_1}, \ldots, \eta_{B_k})$ and $\eta_{B_1 \cup \ldots \cup B_k}$.

Further Results. By looking at Fig. 8 and 9, one can already obtain a global understanding of optimal measurement in the complete graph: As one increases the complexity level (by increasing the trade-off parameter β), there is a transition between the empty measurement (minimal complexity) and the macroscopic measurement (maximal predictive capacity). This transition would be quasi-continuous for a large system, as measurements of intermediate sizes become optimal for different "layers" of the parameter space. These measurements can actually be interpreted as measurements on samples of the agent set, controlling the complexity by adjusting the sample size.

In the following, we provide more formal results regarding the efficiency of feasible measurements in the complete graph.

Observation 6. *For any horizon, as time goes by, the complexity and the predictive capacity of any non-empty measurement $\eta_{\mathcal{A}} = (\eta_{A_1}, \ldots, \eta_{A_k})$ converge to those of a fair Bernoulli variable predicting itself:*

$$\begin{aligned} \forall \tau \in \mathbb{N}, \quad & I(X^t; \eta_{\mathcal{A}}(X^t)) \xrightarrow[t\to\infty]{} 1 \,\text{bit}\,, \\ & I(\eta_{\mathcal{A}}(X^t); \eta_{\Omega}(X^{t+\tau})) \xrightarrow[t\to\infty]{} 1 \,\text{bit}\,. \end{aligned}$$

This result is illustrated in the first plot of Fig. 8 by the fact that the data points of all non-empty measurements converge to the same data point (1 bit, 1 bit).

Proof. Since the system converges to one final state among two possible final states $(0, \ldots, 0)$ or $(1, \ldots, 1)$, the distribution of the state space $\Pr(X^t)$ converges to a Bernoulli distribution. Moreover, since the initial distribution $\Pr(X^0)$ and the transition matrix $T(X^{t+1}|X^t)$ in the case of the complete graph are perfectly symmetric regarding these two possible final states, they are both equally likely. Hence, any non-empty measurement also converge to one of the two possible final states with equal probability. □

Observation 7. *At any time, as the horizon increases, the predictive capacity of any non-empty measurement converges to a non-null value:*

$$\forall t \in \mathbb{N}, \quad \lim_{\tau\to\infty} I(\eta_{\mathcal{A}}(X^t); \eta_{\Omega}(X^{t+\tau})) \;>\; 0.$$

This result is illustrated in the first plot of Fig. 9 by the fact that the data points of all non-empty measurements are above 0 bit.

Proof. Any non-empty measurement provides information about the current microscopic state ($I(\eta_{\mathcal{A}}(X^t); X^t) > 0$) and, henceforth, it conveys information about the following steps ($I(\eta_{\mathcal{A}}(X^t); X^{t+\tau}) > 0$). This is because, in the complete graph, each agent has a potential impact on the system's next global state. □

Observation 8. *At any time and for any horizon, the β-border between the optimality region of the empty measurement and the optimality region of any non-empty measurement is finite. This border converges to 1 from above as time goes by, and to a finite value from below as the horizon increases:*

$$\begin{aligned} \forall t \in \mathbb{N},\ \forall \tau \in \mathbb{N}, \quad & \beta^{t,\tau}_{\emptyset,\mathcal{A}} > 1, \\ \forall \tau \in \mathbb{N}, \quad & \beta^{t,\tau}_{\emptyset,\mathcal{A}} \xrightarrow[t\to\infty]{} \beta^{\infty,\tau}_{\emptyset,\mathcal{A}} = 1, \\ \forall t \in \mathbb{N}, \quad & \beta^{t,\tau}_{\emptyset,\mathcal{A}} \xrightarrow[\tau\to\infty]{} \beta^{t,\infty}_{\emptyset,\mathcal{A}} > 1, \\ & \textit{with} \quad \beta^{t,\tau}_{\emptyset,\mathcal{A}} = \frac{I(X^t; \eta_{\mathcal{A}}(X^t))}{I(\eta_{\mathcal{A}}(X^t); \eta_{\Omega}(X^{t+\tau}))}. \end{aligned}$$

Hence, $\forall \beta \leq 1$, the empty measurement is always more efficient than any non-empty measurement and, $\forall \beta > 1$, there is a time after which any non-empty measurement

becomes more efficient than the empty measurement. Moreover, given a non-empty measurement $\eta_{\mathcal{A}}$, for any time $t \in \mathbb{N}$, if $\beta \in [0, \beta^{t,\infty}_{\emptyset,\mathcal{A}}[$, then the empty measurement is more efficient than the non-empty measurement to predict the final state ($\tau \to \infty$) and, if $\beta \in]\beta^{t,\infty}_{\emptyset,\mathcal{A}}, +\infty[$, then the non-empty measurement is more efficient than the empty measurement to predict the final state.

This result is illustrated in the second plot of Fig. 8 by the fact that the border between the empty measurement and the non-empty measurement of size 1 converges to 1 when $t \to \infty$ and, in the second plot of Fig. 9, by the fact that this border converges to a finite value of the trade-off parameter when $\tau \to \infty$.

Proof. Since the complexity and the predictive capacity of the empty measurement are always null, and since

$$I(\eta_{\mathcal{A}}(X^t); \eta_{\Omega}(X^{t+\tau})) \;=\; H(\eta_{\mathcal{A}}(X^t)) - H(\eta_{\mathcal{A}}(X^t)|\eta_{\Omega}(X^{t+\tau}))$$

with $H(\eta_{\mathcal{A}}(X^t)|\eta_{\Omega}(X^{t+\tau})) > 0$, we have

$$\beta^{t,\tau}_{\emptyset,\mathcal{A}} = \frac{H(\eta_{\mathcal{A}}(X^t))}{I(\eta_{\mathcal{A}}(X^t); \eta_{\Omega}(X^{t+\tau}))} > 1.$$

Moreover, $H(\eta_{\mathcal{A}}(X^t))$ and $I(\eta_{\mathcal{A}}(X^t); \eta_{\Omega}(X^{t+\tau}))$ converge to 1 when $t \to \infty$ (see Observation 6), hence $\beta^{t,\tau}_{\emptyset,\mathcal{A}} \xrightarrow[t\to\infty]{} 1$. Moreover, $H(\eta_{\mathcal{A}}(X^t))$ does not depend on τ and $I(\eta_{\mathcal{A}}(X^t); \eta_{\Omega}(X^{t+\tau}))$ converges to a non-null value when $\tau \to \infty$ (see Observation 7). Hence, $\beta^{t,\tau}_{\emptyset,\mathcal{A}} \xrightarrow[\tau\to\infty]{} \beta^{t,\infty}_{\emptyset,\mathcal{A}} > 1$. □

5 Multilevel Prediction of the Voter Model

This section presents slightly more complex settings of the VM where optimal predictors might be *multilevel* predictors, that is, the joint measurement of agent subsets of different sizes. We show that the need for such multilevel measurements appears in two cases. First, when one wants to predict the state of a subpart of the system. In this case, the subpart's local dynamics and the system's global dynamics compete to determinate future trajectories. Second, the introduction of heterogeneity in the interaction graph might also be responsible for the emergence of several competing dynamics. Note that, in the following experiments, we do not provide any formal proof, but mainly focus on the computed IB-diagrams and their interpretation in terms of multilevel prediction.

5.1 Predicting the State of an Agent in the Complete Graph

Our first results concern the prediction of the agent measurement $\eta_{\{1\}}$ in the complete graph[4], that is the state $X_1^{t+\tau}$ of agent 1 at time $t + \tau$. By applying once

[4]Since the initial state is uniformly distributed and the interaction graph is uniform, the following results do not depend on the chosen agent $i \in \Omega$.

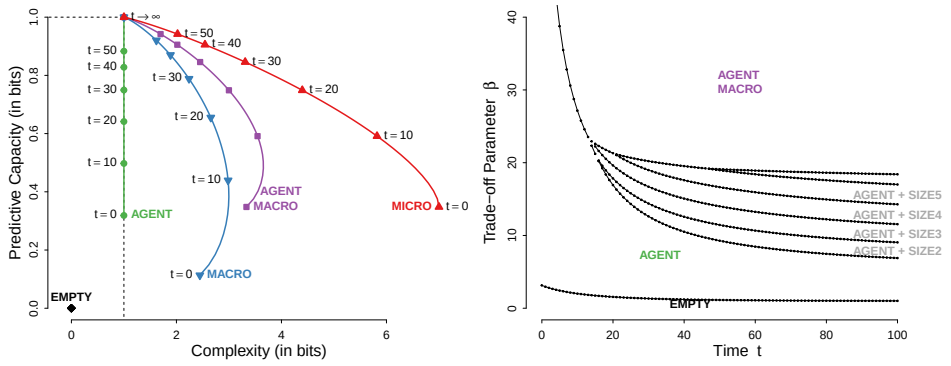


Figure 10: Predicting the state of agent 1 in the complete graph (size $N = 7$, horizon $\tau = 3$, and variable time $t \in \mathbb{N}$)

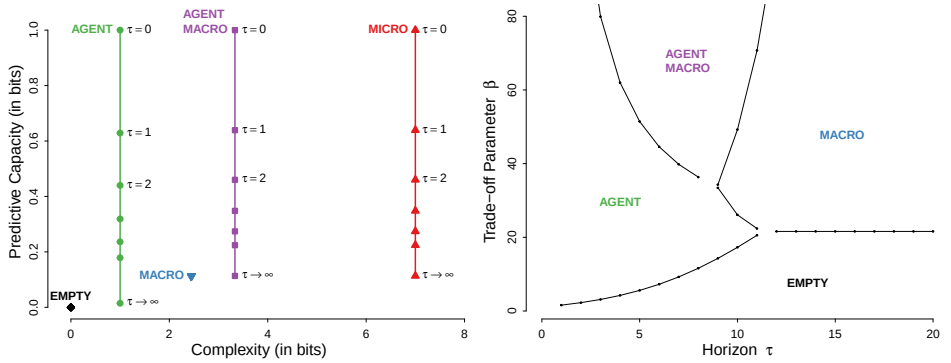


Figure 11: Predicting the state of agent 1 in the complete graph (size $N = 7$, time $t = 0$, and variable horizon $\tau \in \mathbb{N}$)

again the general result of Subsection 3.2 to the subset collection $\{\{1\}, \{2, \ldots, N\}\}$, one could easily show that Theorems 2, 3 and 4 also holds in this setting. Hence, we focus on the following measurements to solve the OPP:

EMPTY	The empty measurement $\eta_\emptyset$;
AGENT	The measurement $\eta_{\{1\}}$ of agent 1;
SIZE2	The measurement of a subset of two agents $\eta_{\{1,i\}}$;
SIZE3	The measurement of a subset of three agents $\eta_{\{1,i,j\}}$;
...	
MACRO	The macroscopic measurement $\eta_\Omega = \eta_{\{1,\ldots,N\}}$;
AGENT+SIZE2	The combination of $\eta_{\{1\}}$ and $\eta_{\{1,i\}}$;

AGENT+SIZE3 The combination of $\eta_{\{1\}}$ and $\eta_{\{1,i,j\}}$;
...
AGENT+MACRO The combination of $\eta_{\{1\}}$ and η_Ω.

In the following, we provide a global understanding of the optimality regions of these measurements based on the plots and IB-diagrams of Fig. 10 and 11 computed for a complete graph of size $N = 7$, with fixed horizon $\tau = 3$ in the case of Fig. 10 and fixed time $t = 0$ in the case of Fig. 11. As shown in Fig. 10, and contrary to the previous case, the macroscopic measurement is never optimal for short-term predictions ($\tau = 3$). On the one hand, for small t, as the system state is likely to be heterogeneous (about as much agents in state 0 than in state 1), the macroscopic measurement conveys very few information regarding the current state of agent 1. However, since according to the transition matrix the probability that agent 1 *does not* synchronise with other agents during the next 3 simulation steps is about 0.63 (so that $\eta_{\{1\}}(X^t) = \eta_{\{1\}}(X^{t+3})$ in more than half of the cases), the current agent state does provide some useful information for short-term prediction. If, to the contrary, agent 1 *does* synchronise with another agent, because the system state is heterogeneous, the result of this synchronisation is very difficult to predict and knowing the aggregated global state $\eta_\Omega(X^t)$ does not help much. Hence, for small t, the agent measurement $\eta_{\{1\}}(X^t)$ is more predictive (and obviously less complex) than the macroscopic measurement $\eta_\Omega(X^t)$.

On the other hand, for larger t, as the system state becomes more homogeneous (many agents in state 0 or many agents in state 1), the macroscopic measurement becomes more efficient because (1) it conveys information about the agent current state and (2) the result of a synchronisation of agent 1 with any other agent becomes easier to predict from the system's global state. Hence, the macroscopic measurement becomes more predictive than the agent measurement. However, in this context, the multilevel measurement $(\eta_{\{1\}}, \eta_\Omega)(X^t)$, mixing the agent state and the global state, consists in an even more efficient short-term predictor. This is because adding the agent measurement to the macroscopic measurement does not increase much its complexity (especially for homogeneous states), but significantly increases its predictive capacity by taking into account the cases where agent 1 is not in the majority and does not synchronise during the 3 simulation steps. Moreover, this multilevel measurement becomes more efficient as time goes by (optimal for smaller β when t increases), because the macroscopic measurement becomes itself less complex and more predictive.

As shown in Fig. 11, three non-empty measurements might be interesting to predict the agent state depending on the prediction horizon and the allowed complexity level. If one is interested in short-term prediction (small τ), as we said previously, the agent measurement is quite adequate since it contains only local information regarding the agent current state (that is likely to stay the same in the short-term). For long-term predictions (large τ), the macroscopic measurement becomes the best predictor because the system's final state – and hence the agent's final state – depends more on the initial global state than on the initial agent state. More interestingly, the multilevel measurement is mostly adequate for intermediate-

term prediction. This is because this measurement provides an interesting mixture when both the agent local state and the system global state are likely to influence the agent future state. Hence, this example shows how the need for a multilevel prediction is really related to the temporality of prediction. It arises when the local and global dynamics mix and generate what we can consequently call *multilevel dynamics*.

5.2 Impact of Heterogeneous Interaction Patterns on Prediction Efficiency

In the following experiments, we introduce some heterogeneity within the interaction graph of the VM. To do so, we examine three cases of the following family of models, that we call the two-community Voter Models: The agent set Ω is partitioned into two disjoint groups Ω_1 and Ω_2 – or *communities*, which both consist in a complete and uniform interaction graph, and such that the interaction patterns between agents of each community are also complete and uniform (see Fig 12). Hence, this family of models depends on 6 parameters:

- $N_1 \in \mathbb{N}$ and $N_2 \in \mathbb{N}$ are the number of agents in each community;
- $\rho_{1\leftrightarrow 1} \in \mathbb{R}^+$ and $\rho_{2\leftrightarrow 2} \in \mathbb{R}^+$ are the weights of edges between agents of the same community;
- $\rho_{1\to 2} \in \mathbb{R}^+$ and $\rho_{2\to 1} \in \mathbb{R}^+$ are the weights of edges between agents of different communities.

These weights are simply used to compute the probability of choosing a given edge at each simulation step as the ratio between the weight of the edge and the sum of all weights. The complete VM is hence a member of this more general family for which all weights are equal ($\rho_{1\leftrightarrow 1} = \rho_{2\leftrightarrow 2} = \rho_{1\to 2} = \rho_{2\to 1}$). In the following, we present results for community sizes $N_1 = 10$ and $N_2 = 10$, for intra-community weights $\rho_{1\leftrightarrow 1} = 1$ and $\rho_{2\leftrightarrow 2} = 1$, but for variable inter-community weights $\rho_{1\to 2}$ and $\rho_{2\to 1}$. Moreover, we are still interested in the prediction of $\eta_{\{1\}}$, that is the state of agent 1, which belongs in community 1.

Once again, by showing that the measurement resulting from the subset collection $\{\{1\}, \Omega_1, \Omega\}$ (see bottom-left drawing in Fig 12) is fully informative of future states of agent 1 and has the uniform micro-state property, one could apply the general result of Subsection 3.2 and build a list of potentially optimal measurements. However, for simplicity, we focus in the following on a subset of this list (see Fig 12 for a graphical representation of some of them):

EMPTY	The empty measurement $\eta_\emptyset$;
AGENT	The measurement $\eta_{\{1\}}$ of agent 1;
MESO1	The measurement η_{Ω_1} of the aggregated state of agents in community 1;
MESO2	The measurement η_{Ω_2} of the aggregated state of agents in community 2;
MACRO	The macroscopic measurement $\eta_\Omega = \eta_{\{1,\dots,N\}}$;

and the five possible combinations: **(AGENT + MESO1)**, **(AGENT + MESO2)**, **(AGENT + MACRO)**, **(MESO1 + MACRO)**, and **(AGENT + MESO1 + MACRO)**. Note that, because η is additive, other possible combinations are actually equivalent with these ones (see Observation 4): *E.g.*, **(MESO1 + MACRO)**, **(MESO2 + MACRO)** and **(MESO1 + MESO2)** are equivalent. The Hasse diagram of the corresponding lattice structure is presented in Fig. 13 (see also Subsection 3.1).

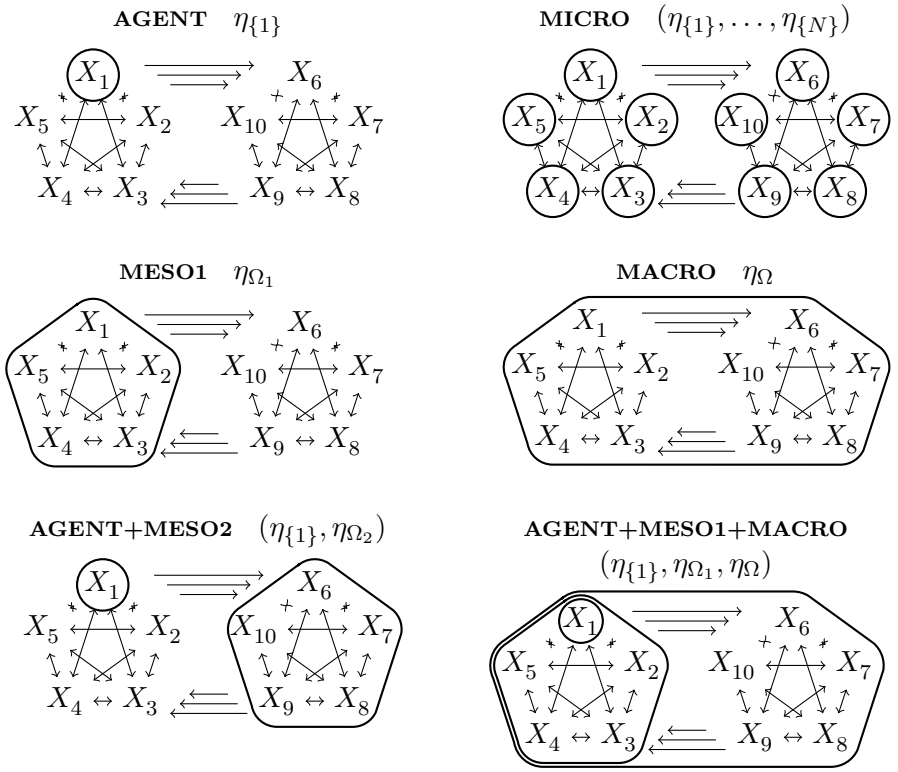


Figure 12: Example feasible measurements defined on the two-community Voter Model ($N_1 = 5$ and $N_2 = 5$)

5.2.1 The Symmetric Two-community Case

The first experiment deals with what is classically understood by a "community structure". The system is made up of two sets of agents which are more likely to interact with agents of the same set than with agents of the other set (see Fig. 14a). Here, by taking $\rho_{1\to 2} = \frac{1}{5}$ and $\rho_{2\to 1} = \frac{1}{5}$, an edge between two agents of the same community is 5 times more likely to be chosen at each simulation step than an edge between agents of different communities.

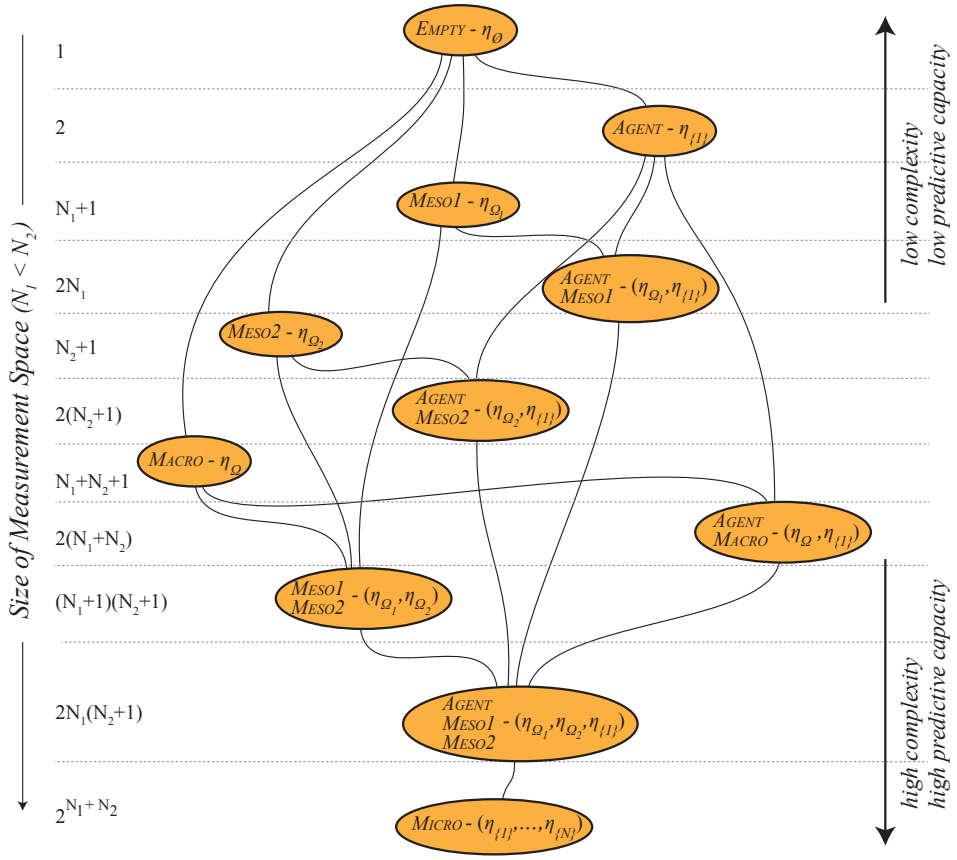


Figure 13: Hasse diagram of the poset structure of canonical feasible measurements in the two-community Voter Model

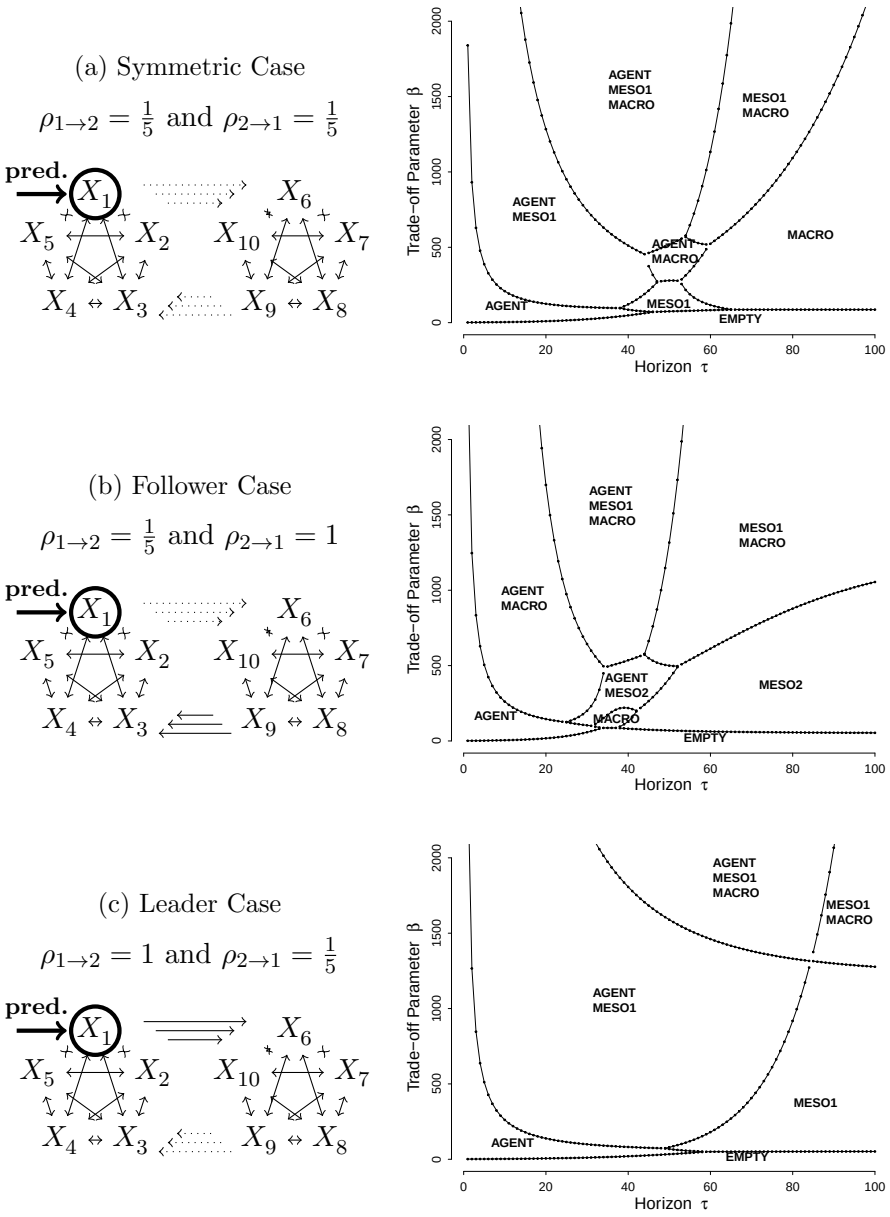


Figure 14: Predicting the state of agent 1 in three different settings of the two-community voter model (sizes $N_1 = 10$ and $N_2 = 10$, inter-community weights $\rho_{1\leftrightarrow1} = 1$ and $\rho_{2\leftrightarrow2} = 1$, time $t = 0$, and variable horizon $\tau \in \mathbb{N}$)

As a result of this heterogeneous structure, the synchronisation process within a community is expected to be quicker than the synchronisation of the whole system. When predicting the state of a particular agent, this implies three levels of dynamics with different temporal scales: (1) the agent state at a microscopic level has a short-term influence, (2) the community state at a mesoscopic level has a intermediate-term influence, and (3) the system state at a macroscopic level has a long-term influence. These three levels are indeed observed in the IB-diagram of Fig. 14a where the agent measurement $\eta_{\{1\}}$, the mesoscopic measurement η_{Ω_1} and the macroscopic measurement η_Ω are successively optimal for small values of β as the horizon τ increases. As in the complete graph setting, the prediction of such multilevel dynamics might also requires multilevel predictors when one allows a higher complexity (larger values of β): (4) the agent and mesoscopic measurements can be combined $(\eta_{\{1\}}, \eta_{\Omega_1})$ for more efficient short-term prediction, (5) the mesoscopic and macroscopic measurements can be combined $(\eta_{\Omega_1}, \eta_\Omega)$ for more efficient long-term prediction, and (6) combining the three levels $(\eta_{\{1\}}, \eta_{\Omega_1}, \eta_\Omega)$ also becomes optimal for large values of β in the case of intermediate-term prediction.

Interestingly, there are also cases where the microscopic and macroscopic levels should be combined for optimal prediction $(\eta_{\{1\}}, \eta_\Omega)$, without taking into account the intermediate mesoscopic level. This result for intermediate-term prediction and intermediate complexity level is explained by distinguishing two cases. First, in the case the agents in community 1 are highly heterogeneous at time $t = 0$, knowing the initial mesoscopic state is not very relevant for prediction. To the contrary, predicting the future state of agent 1 significantly benefits from the knowledge of the initial macroscopic state. Second, in the case the agents in community 1 are more homogeneous at time $t = 0$, the initial mesoscopic state becomes highly relevant for prediction. But, in this case, knowing only the initial state of agent 1 is often sufficient to know the current mesoscopic state. Hence, the agent-macroscopic measurement $(\eta_{\{1\}}, \eta_\Omega)$ is quite efficient for intermediate-term prediction, that is when the state of community 1 is likely to have converged, because it takes into account the two different cases, contrary to the mesoscopic measurement η_{Ω_1} that does not.

5.2.2 The Follower Case

In this second experiment, only the weights of the edges from community 2 to community 1 are reduced ($\rho_{2\to1} = \frac{1}{5}$ and $\rho_{1\to2} = 1$). We refer to this case as the *follower* case, as opposed to the *leader* case addressed in the next subsection, since agent 1 is 5 times more likely to be influenced by an agent of community 2 than to influence such an agent in return.

The IB-diagram obtained with this setting (Fig. 14b) is qualitatively comparable with the one of the symmetric two-community case (Fig. 14a). We indeed rediscover 8 regions organised according to the prediction horizon and the allowed complexity level. However, we notice two significant changes:

1. The mesoscopic measurement η_{Ω_1} of community 1 in Fig. 14a is systematically replaced by the macroscopic measurement η_Ω in Fig. 14b. This is not

surprising since, in the follower case, and contrary to the two-community case, agent 1 is as likely to be influenced by agents of community 2 than by agents of community 1. Hence, for intermediate-term prediction, one can efficiently take into account the macroscopic state (instead of the mesoscopic state) to predict the future state of agent 1.

2. The macroscopic measurement η_Ω in Fig. 14a is systematically replaced by the mesoscopic measurement η_{Ω_2} of community 2 in Fig. 14b. This is explained by the fact that the dynamics of the whole system are now mainly steered by community 2. Indeed, agents of community 1 are little likely to influence the global state on the long-term. Hence, one can efficiently avoid to measure the global state of community 1 for long-term prediction and focus on community 1.

This experiment shows that the ordering of levels when we talk about "multi-level prediction" is not necessarily related to the size of the measured agent subsets, but rather to the temporality of the interaction processes within these subsets. In this example, the long-term dynamics, that will determine the system's final state, are better predicted by the current state of a "small" agent subset Ω_2 that has a high influence on the long-term, whereas the intermediate-term dynamics are better predicted by the current state of the whole agent set Ω which influence is actually restricted in time. Hence, because we deal with complex heterogeneous dynamics, "high-level" measurement does not necessarily mean "many-agents" measurement.

5.2.3 The Leader Case

In this third experiment, only the weights of the edges from community 1 to community 2 are reduced ($\rho_{1\to 2} = \frac{1}{5}$ and $\rho_{2\to 1} = 1$). Hence, agent 1 is five times more likely to influence an agent of community 2 than to be influenced by such an agent in return.

In this last case, the IB-diagram in Fig. 14c can be summarised by three cuts of the parameter space inducing six optimality regions. First, there is one "vertical" cut between short-term prediction (on the left of the diagram) where the optimal measurements always include the current state of agent 1, and long-term prediction (on the right of the diagram) where the current state of agent 1 is not taken into account. This "vertical" cut is thus induced by the decreasing predictive capacity of the agent measurement $\eta_{\{1\}}$ as the horizon increases. Second, there are two "horizontal" cuts delimiting 3 prediction levels respectively characterised by the empty measurement $\eta_\emptyset$, the mesoscopic measurement η_{Ω_1}, and the multilevel measurement $(\eta_{\Omega_1}, \eta_\Omega)$. These "horizontal" cuts thus correspond to the complexity levels that might be allowed for prediction by successively adding measurements. Indeed, in this case, the macroscopic measurement η_Ω is never individually optimal because it does not take into account the heterogeneous interaction structure. Therefore, the only optimal choice is here to avoid macroscopic observation, or to combine it with mesoscopic observation. To conclude, in this last case, the IB-diagram is

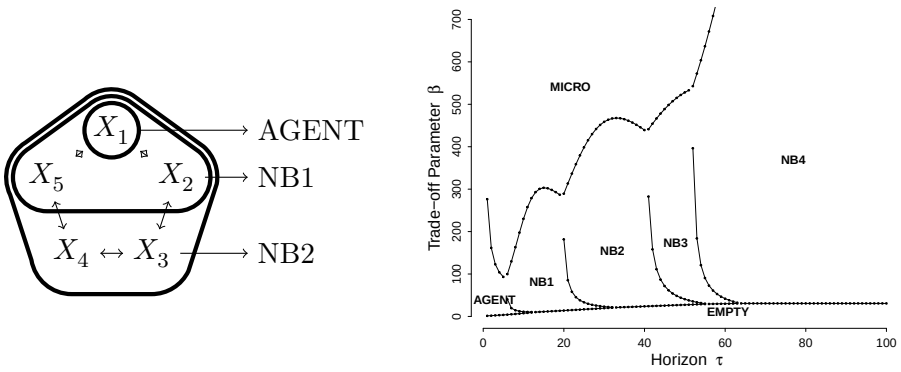


Figure 15: Predicting the state of agent 1 in the ring (size $N = 9$, time $t = 0$, and variable horizon $\tau \in \mathbb{N}$)

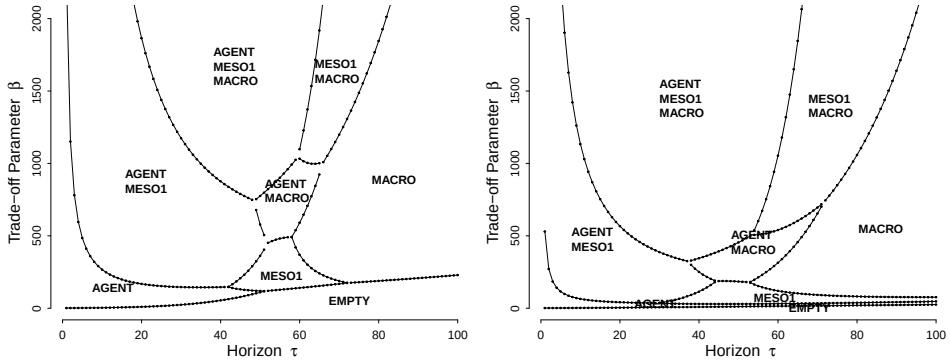
partitioned into 2×3 regions that depend both on the prediction horizon and on the trade-off parameter.

5.2.4 Optimal Predictor Size in The Ring

In this last experiment, we consider yet another king of interaction graph: a *ring* (see Fig. 15). Each agent is able to directly interact with exactly two neighbours, but it is also globally connected to any other agent from edge to edge in a circular chain.

The objective here is to predict the state of agent 1 by measuring the current aggregated state of its close or distant neighbourhoods. For example, in Fig. 15, "NBx" designs the "neighbourhood of size x", that is the aggregated state of agent 1 and of all agents that it can reach by x edges in the ring. In this setting, we expect that, for any given horizon $\tau \in \mathbb{N}$, there is an optimal neighbourhood size that one should use for prediction. This means in practice that, regardless of the allowed complexity level, and hence regardless of the value of the trade-off parameter, one neighbourhood size is always preferred to the others for local prediction. Moreover, we expect that the optimal neighbourhood size is small for short-term prediction, as including the state of far agents does not provide much information to predict the state of agent 1 in the near future, and that, to the contrary, it is large for long-term prediction, as all agents might participate to the system's convergence toward its final state.

These expectations are confirmed in the IB-diagram of Fig. 15 by relatively sharp vertical transitions between the optimality regions of neighbourhood measurements as the horizon increases. Hence, as a result, the optimal neighbourhood size does not depend much on the trade-off parameter, and one unique size is mainly associated to each horizon regardless of the allowed complexity level.



Transient Dynamics ($t = 0$) Stationary Dynamics ($t = +\infty$)

Figure 16: Contrarian Case: Predicting the state of agent 1 in the contrarian Voter Model (sizes $N_1 = 10$ and $N_2 = 10$, inter- and intra-community weights $\rho_{1\leftrightarrow 1} = 1$, $\rho_{2\leftrightarrow 2} = 1$, $\rho_{1\to 2} = \frac{1}{5}$, and $\rho_{2\to 1} = \frac{1}{5}$, contrarian probability $q = \frac{1}{21}$, and variable horizon $\tau \in \mathbb{N}$)

5.3 The Contrarian Case

The *contrarian* VM slightly generalises the classical VM by allowing the agents to spontaneously desynchronise with their neighbours: given a contrarian probability $q \in [0, 1]$, at each simulation step – when a directed edge (i, j) is randomly selected – agent j might take the opposite state of agent i with probability q, instead of taking the same state as in the classical VM (which corresponds to the $q = 0$ case). In the context of opinion dynamics, contrarian behaviour relates to the presence of individuals that do not seek conformity in all cases or to the existence of certain situations in which agents would not desire to adopt the opinion of their interaction partner. It has been introduced into majority rule opinion models [13] and here we use the contrarian version of the VM which has been previously considered in [1]. Hence, given a directed edge (i, j), we have:

$$\Pr(X_j^{t+1} \neq X_i^t) = q, \quad \Pr(X_j^{t+1} = X_i^t) = 1 - q, \quad \text{and } \forall k \neq i,\ \Pr(X_k^{t+1} = X_k^t) = 1.$$

For $q > 0$ this induces a dynamics with no absorbing states as the synchronised configurations with all agents in the same state are left with probability q. It rather leads to a Markov chain with a well-defined stationary distribution showing an ordered phase with switching between the two consensus states for $q < 1/(N + 1)$ and a disordered phase for larger q. For the complete graph the equilibrium distribution is uniform for $q^* = 1/(N + 1)$ (see [1] for details), and we used this contrarian rate in our analyses.

We now consider two prediction problems where in both cases the aim is to

predict the state of a single agent. In the first case, we start at time $t = 0$ and try to anticipate the system's transient dynamic, and in the second case we consider the prediction problem starting at time $t = \infty$ when the system is in the stationary regime. Fig. 16 gives the result of these two versions of the prediction problem applied to the agent measurement $\eta_{\{1\}}$ in the symmetric two-community case with 10 agents in each population and inter- and intra-community weights $\rho_{1\leftrightarrow 1} = 1$, $\rho_{2\leftrightarrow 2} = 1$, $\rho_{1\to 2} = \frac{1}{5}$, and $\rho_{2\to 1} = \frac{1}{5}$ (see Subsection 5.2.1). In this case, the contrarian probability q^* is $\frac{1}{21}$.

A first general observation, when going from the absorbing VM to the non-absorbing contrarian case, is that the empty measurement becomes optimal as $\tau \to \infty$ in the contrarian case. This is due to the fact that the dynamics is more and more uncorrelated and the mutual information $I(X^t; X^{t+\tau})$ vanishes as τ increases. Therefore, any pre-measurement $\eta_A(X^t)$ will eventually loose its predictive capacity too such that the empty measurement – being the least complex one – is optimal from the IB point of view.

Despite this difference, the structure of the IB-diagrams shown in Fig. 16 is fairly similar to the corresponding diagram for the classical VM in Fig. 14a. This is especially true for the transient dynamics (notice that $0 \leq \beta \leq 2000$ in Fig. 14a whereas $0 \leq \beta \leq 1000$ in Fig. 16). One interesting observation for the stationary dynamics is that the measurement of the agent to be predicted (**AGENT**) and the measurement of the community to be predicted (**MESO1**) both stay optimal for a small range of relatively low β (low complexity) even when the prediction horizon increases and exceeds $\tau = 100$. This effect is neither observed in the absorbing case nor in the transient regime of the contrarian VM.

6 Conclusion and Perspectives

6.1 Summary

This paper presents three main contributions regarding the general problem of efficient prediction of dynamical systems. First, it proposes a generic formalism for the concept of "prediction efficiency" through a constrained optimisation problem: the *Optimal Prediction Problem* (OPP). More precisely, we propose to use the Information Bottleneck (IB) framework to model the two competitive objectives of a prediction task: (1) minimising the complexity of the measurement that is used for prediction and (2) maximising its predictive power. We also define a solution to the OPP as a tridimensional diagram representing the optimality regions of pre-measurements in the parameter space (current time, prediction horizon, and trade-off parameter). This framework seems to be quite powerful to define a sound and generic setting from an information-theoretic perspective, that is independent of particular considerations that would emerge from practical applications (such as the real *cost* of measurement or the *reward* of proper prediction). However, as described below in Subsection 6.3, we claim that our approach could be generalised to other (more sophisticated) objective functions.

Our second contribution is a detailed analysis of the OPP solution space in the case of the prediction of Agent-based Models (ABMs). To this purpose, we introduce the concept of *generic measurement* – that is a measurement that can be generically applied to any agent subset, hence defining combinatorial constraints on the solution space. We then formally describe the algebraic structure of this constrained solution space by providing general combination rules for *additive* measurements. We show how this structure can be exploited to solve the OPP and, in particular, to reduce its computational complexity. These contributions can be considered as the core of our theoretical work and we hope to generalise it to other objective functions (assuming monotonicity with respect to the *refinement relation*, see Subsection 3.1) and to other settings (such as the problem of *level identification*, see Introduction).

Third, we apply this theoretical framework to a classical and well-known ABM: the *Voter Model* (VM). This allows us to show how our theoretical contributions should be used in practice, but also to provide results that could be generalised to other *diffusion processes* defined under similar hypotheses. Here is a summary of the three main results of these experiments:

1. The microscopic level is not more informative than the aggregated level when predicting macroscopic observables of homogeneous systems, that is when the agents in each aggregate contribute similarly to the dynamics of the macroscopic state. In case of slightly heterogeneous behaviour, this result could also be used as a heuristic to find efficient observation levels.

2. When predicting the state of some subpart of the system, optimality strongly depends on the prediction horizon: local measurements are more efficient for short-term prediction, global measurements for long-term prediction, and multilevel measurements for intermediate-term prediction, that is when the system's dynamics can be efficiently described as a mixture of local and global processes.

3. Heterogeneity in the agent behaviour might require the observer to refine the description levels and, in particular, to introduce mesoscopic levels that takes into account the system's internal structure. In the future, we hope to formalise this result in a more systematic way in order to provide general rules to decompose a system in relevant scales given is structure.

6.2 Application Perspectives

In this paper we studied multilevel prediction only for a simple theoretical model. An important issue for the future will concern the application to more realistic situations. We envision two interrelated ways of how this could be achieved. On the one hand, the analysis of efficient prediction in a real complex system could be based on models that more accurately describe the system at question but still allow to derive a Markovian microscopic transition kernel along with the specification of the state space partitions induced by a set of measurements and in which it is

therefore possible to compute the involved information measures. On the other hand, we can also aim at adapting the framework in such a way that we can apply it to real data, in which case, usually, no complete knowledge of the micro dynamics is given and therefore model inference should be reflected in the cost term too.

The VM has been originally introduced as a model of spatial conflict of different species [19]. The framework we propose could be applicable in the field of ecology as a way to provide theoretical support for decisions concerning the measurement and data collection stage (pre-measurements) with the aim to identify observables that contain information for a particular prediction purpose (post-measurement) and reduce at the same time the data acquisition costs. For instance, if the aim is to predict the presence of a certain species at a specific site (agent measurement), our experiments indicate that in the short run data collected at this site might be the most useful one. However, depending on the prediction horizon, additional data – regional or even global – can complement this measure and add important information about the prediction problem to be addressed. The notion of multiple levels may have a two-fold meaning here, one related to the structure of the geographical space and the other one to the structure of inter-species mutual dependencies [22, 26, 27]. Both, the history of the species at question at different geographical scales as well as its embeddedness into the webs of food and reproduction could be relevant and the framework, applied to a more sophisticated model that includes these aspects, could provide an idea which observables may contain useful additional information.

To make another example, in recent years, the UN statistical department as well as other agencies have made available large amounts of data on the trade of different products between the countries of the world[5]. On the basis of these data, measures of economic complexity [17] and fitness [29] have been constructed by aggregating information from the structure of the exports of countries into a single observable. These measures were shown to have significant predictive power to anticipate the growth potential of countries. The Information Bottleneck framework can be a useful tool for the assessment of the predictive capacity of these measures and to evaluate other observables at different aggregation levels as well as combinations of them. In this context, as in the previous example, macroscopic observables can be defined along at least two dimensions: aggregation of trade data in the geographic space (from regional trade to international trade) and in the product space (from a refined list of products to aggregated classes such as industrial sectors or production chains). This application will be addressed in a forthcoming paper [3].

6.3 Theoretical Perspectives

Currently, a fundamental hypothesis of our framework is that a complete model of the system's microscopic dynamics is available to the observer. However, in most

[5]The UN statistical department makes this data available under the COMTRADE web page (`http://comtrade.un.org`) and the French research centre in international economics CEPII (`http://www.cepii.fr`) provides two rectified datasets on international trade partly based on COMTRADE data.

practical cases – such as in the two application fields here above mentioned – no such model exists. Hence, the predictor – that is the conditional probability of the post-measurement given the pre-measurement – has to be inferred from a limited amount of data. If one or both measurements are low-level observables this can become very challenging – or will be not feasible at all – because they are usually very high-dimensional and the data requirements exponentially increase, making inference at low levels very often unfeasible in practice. In order to deal with this major issue one has to take into account the degree of inferability of predictors as part of the cost function of the optimisation problem. Hence, efficient prediction would be driven in this context not only by the complexity of the measurements alone but also, to a large extent, by the complexity of the predictor, *i.e.* the *model complexity*. If we start with a parametrized model for the conditional probability on the lowest level (given by the data), the refinement relation on the pre-measurement will also induce a hierarchy on the induced models. Thus, the corresponding optimisation problem is directly related to the problems of over-fitting and model selection in statistical inference. For instance, in the case of maximum likelihood estimation regularisation terms such as the Akaike Information Criterion (AIC) would take the part of the model complexity, see also [15] for a Bayesian perspective. In general, we propose to rely on classical work in learning theory, such as the theoretical tools of model selection and feature selection to express the trade-off between the model likelihood (quantifying how well the estimated macro-dynamics fit with the empirical data) and the model complexity (that should be controlled to avoid over-fitting at microscopic levels).

Moreover, in many real-world scenarios, prediction is strongly related to financial cost of data acquisition and economic impact of prediction-based decisions. For example, one would like to optimise investments based on the prediction of future economic indicators with limited money allocated to data collection. To be useful in this context, objective functions should express domain-depend costs and rewards. We hence plan to generalise our framework to a broad class of such objectives, thus going over the IB framework. For example, a cost could be associated to any outcome of the pre-measurement and a reward to any outcome of the post-measurement. The OPP would consist in maximising the expected reward while minimising the expected cost. To give an example, when dealing with extreme events and crisis prediction, outliers are much more important to predict than regular trajectories of the system. The resulting objective function can hence be highly non-regular, but yet satisfy general properties that are required and sufficient for our framework to apply. To this purpose, one has yet to make this properties explicit – such as the monotonicity of objective functions with respect to the refinement relation – and to show which classes of objectives can hence be optimised with our method.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement no. 318723 (MatheMACS). S.B. also acknowledges financial support by the Klaus Tschira Foundation.

References

[1] S. Banisch. From Microscopic Heterogeneity to Macroscopic Complexity in the Contrarian Voter Model. *Advances in Complex Systems*, 17:1450025, 2014.

[2] S. Banisch. *Markov Chain Aggregation for Agent-Based Models*. Understanding Complex Systems. Springer, 2015, forthcoming. Preliminary Version available at `http://pub.uni-bielefeld.de/publication/2690117`.

[3] S. Banisch, R. Lamarche-Perrin, and E. Olbrich. Evaluating Multilevel Predictions from Data – The Case of Trading Data to Predict GDP Growth. In *Proceedings of the 2015 Conferences on Complex Systems (CCS'15)*, 2015, forthcoming.

[4] S. Banisch and R. Lima. Markov Chain Aggregation for Simple Agent-Based Models on Symmetric Networks: The Voter Model. *Advances in Complex Systems*, 2015, in press. arxiv.org/abs/1209.3902.

[5] S. Banisch, R. Lima, and T. Araújo. Agent Based Models and Opinion Dynamics as Markov Chains. *Social Networks*, 34:549–561, 2012.

[6] D. Blackwell. Equivalent Comparisons of Experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953.

[7] C. Castellano, S. Fortunato, and V. Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591–646, 2009.

[8] P. Clifford and A. Sudbury. A model for spatial conflict. *Biometrika*, 60(3):581–588, 1973.

[9] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.

[10] T. M. Cover and J. A. Thomas. Rate Distortion Theory. In *Elements of Information Theory*, pages 336–373. John Wiley & Sons, Inc., 1991.

[11] J. P. Crutchfield and K. Young. Inferring statisitical complexity. *Physical Review Letters*, 63(2):105–108, 1989.

[12] B. A. Davey and H. A. Priestley. *Introduction to Lattices and Order*. Cambridge University Press, Cambridge, UK, Seconde Edition edition, 2002.

[13] S. Galam. Contrarian deterministic effects on opinion dynamics: "the hung elections scenario". *Physica A: Statistical Mechanics and its Applications*, 333(C):453–460, 2004.

[14] D. Geiger, T. Verma, and J. Pearl. Identifying independence in bayesian networks. *Networks*, 20(5):507–534, 1990.

[15] A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.

[16] P. Grassberger. Toward a quantitative theory of self-generated complexity. *International Journal of Theoretical Physics*, 25(9):907–938, 1986.

[17] C. A. Hidalgo and R. Hausmann. The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, 106(26):10570–10575, 2009.

[18] J. G. Kemeny and J. L. Snell. *Finite Markov Chains*. Springer, 2nd ed. edition, 1976.

[19] M. Kimura and G. H. Weiss. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49:561–576, 1964.

[20] R. Lamarche-Perrin. multilevel_prediction. `https://github.com/Lamarche-Perrin/multilevel_prediction/`, 2015.

[21] T. M. Liggett. *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes*, volume 324 of *Grundlehren der mathematischen Wissenschaften*. Springer, 1999.

[22] J. M. Montoya, S. L. Pimm, and R. V. Solé. Ecological networks and their fragility. *Nature*, 442(7100):259–264, 2006.

[23] P. A. P. Moran. Random processes in genetics. In *Proceedings of the Cambridge Philosophical Society*, volume 54, pages 60–71, 1958.

[24] O. Pfante, E. Olbrich, N. Bertschinger, N. Ay, and J. Jost. Closure measures for coarse-graining of the tent map. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 24(1):013136, 2014.

[25] O. Pfante, E. Olbrich, N. Bertschinger, N. Ay, and J. Jost. Comparison between different methods of level identification. *Advances in Complex Systems*, 17(2):1450007, 2014.

[26] S. L. Pimm. *Food Webs*. Springer, 1982.

[27] M. L. Rosenzweig. *Species Diversity in Space and Time*. Cambridge University Press, 1995.

[28] C. R. Shalizi. *Causal Architecture, Complexity and Self-Organization in Time Series and Cellular Automata*. PhD thesis, University of Wisconsin at Madison, Physics Department, 2001.

[29] A. Tacchella, M. Cristelli, G. Caldarelli, A. Gabrielli, and L. Pietronero. A new metrics for countries' fitness and products' complexity. *Scientific reports*, 2, 2012.

[30] N. Tishby, F. C. Pereira, and W. Bialek. The Information Bottleneck Method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing (Allerton'99)*, pages 368–377, September 1999.

[31] D. H. Wolpert, J. A. Grochow, E. Libby, and S. DeDeo. Optimal High-Level Descriptions of Dynamical Systems. *arXiv*, (1409.7403), 2015.

[32] D. Zambella and P. Grassberger. Complexity of Forecasting in a Class of Simple Models. *Complex Systems*, 2:269–303, 1988.