Eur. Phys. J. Spec. Top. (2021) 230:3249–3271
https://doi.org/10.1140/epjs/s11734-021-00191-0

THE EUROPEAN PHYSICAL JOURNAL SPECIAL TOPICS

Regular Article

Statistical analysis of tipping pathways in agent-based models

Luzie Helfmann[1,2,3,a], Jobst Heitzig[3], Péter Koltai[1], Jürgen Kurths[3,4], and Christof Schütte[1,2]

[1] Institute of Mathematics, Freie Universität Berlin, Berlin, Germany
[2] Zuse Institute Berlin, Berlin, Germany
[3] Department of Complexity Science, Potsdam Institute for Climate Impact Research, Potsdam, Germany
[4] Department of Physics, Humboldt University, Berlin, Germany

Received 4 March 2021 / Accepted 31 May 2021 / Published online 18 June 2021

Abstract Agent-based models are a natural choice for modeling complex social systems. In such models simple stochastic interaction rules for a large population of individuals on the microscopic scale can lead to emergent dynamics on the macroscopic scale, for instance a sudden shift of majority opinion or behavior. Here we are introducing a methodology for studying noise-induced tipping between relevant subsets of the agent state space representing characteristic configurations. Due to a large number of interacting individuals, agent-based models are high-dimensional, though usually a lower-dimensional structure of the emerging collective behaviour exists. We therefore apply Diffusion Maps, a non-linear dimension reduction technique, to reveal the intrinsic low-dimensional structure. We characterize the tipping behaviour by means of Transition Path Theory, which helps gaining a statistical understanding of the tipping paths such as their distribution, flux and rate. By systematically studying two agent-based models that exhibit a multitude of tipping pathways and cascading effects, we illustrate the practicability of our approach.

1 Introduction

Understanding tipping pathways and tipping cascades in social systems are very important for our interconnected world. Tipping is defined as a qualitative change from one rather stable state to another one upon a small quantitative change, e.g., of a parameter, or due to noise. One can distinguish between the following general types of tipping [2]: in *bifurcation-induced* tipping, the state of a parameter-dependent system changes qualitatively due to an external control parameter crossing a threshold value. The parameter is assumed to vary infinitesimally slowly such that one studies transitions of stationary dynamics. Often this threshold value is called the *tipping point* or the *point of no return*. In *noise-induced* tipping, noisy fluctuations result in the system escaping from the neighborhood of a metastable state. Last, *rate-induced* tipping happens when the control parameter changes faster than a certain critical rate of change such that the system fails to track the continuously changing attractor.

Social systems are complex systems and as such characterized by rich, nonlinear and usually local (i.e., only between neighbours) interactions among a large number of individual constituents [10,11]. The individual entities are ignorant of the behaviour of the system as a whole and only respond to local information. Social systems are usually open, i.e., continuously in interaction with their environment, and, therefore, often not in a simple equilibrium. Moreover, the history of the system affects the present. Hierarchies and multi-scale structures are present in complex social systems [62]. When modeling social systems, agent-based models (ABMs) are a natural choice. One defines the characteristics of a large sample of discrete entities, the so-called agents (e.g., people, animals, cars, companies, ...), and a set of possible actions and local interactions rules for the agents. Often these are stochastic, thus reflecting the unpredictability and individuality of the agents. From the interplay of local interactions, global patterns can emerge [39].

The complexity of social systems is inherited by their tipping dynamics. In [75], it is argued that tipping in social systems, such as an epidemic, a social contagion process of ideas or norms [20,48], a crash in a financial system, or the diffusion of a new technology [59], faces several difficulties and can be much more complex than tipping in climate or ecological systems. "Tipping in realistic social systems can usually not be linked to a single control parameter, instead multiple interrelated factors act as a forcing of the transitions (e.g. policies, communication, taxation, ...). Moreover, there is a larger number of mechanisms that cause tipping and various pathways of change towards a greater number of potentially stable post-tipping states." [75] Recently, there has been an increasing interest in studying *tipping cascades*, i.e., cascading effects in interacting sys-

[a] e-mail: luzie.helfmann@fu-berlin.de (corresponding author)

tems where the tipping of one sub-system influences the tipping likelihood of another one. Interactions between tipping elements have been studied in the climate system [7,33], in ecological systems [55], and also in social systems [59].

The aim of this paper is two-fold: (i) We propose a methodology how noise-induced tipping in high-dimensional agent-based models can be analysed by combining several existing methods such as nonlinear dimensionality reduction and Transition Path Theory [43,73]. The method can be applied without any restrictive assumptions about the models and only relies on given simulation data. Its efficacy depends implicitly on (possibly unknown) low-dimensional structures in the dynamics. (ii) We demonstrate the applicability of the approach on two typical models of social network-based opinion and behavioural change. The first model describes the threshold-based activation of people for some collective action and is adapted from [21,72] to exhibit noise-induced tipping. We introduce a second model of opinions and possibly differing actual behaviours that displays cyclic tipping.

The idea of our approach is to first reduce the dimension of the model by means of a nonlinear dimension reduction technique, Diffusion Maps, thereby relying on the existence of some lower-dimensional structure on which the ABM dynamics essentially takes place. This assumption can be made for most ABMs, since the emergence of macro-scale patterns and collective behaviour is a key property of ABMs. Central to this is that we do not need to know which macroscopic features the system eventually evolves along, but we can learn the associated coordinates from sufficiently rich dynamical data [30]. Diffusion Maps have already been applied for finding low-dimensional coordinates, so-called *collective variables*, *reaction coordinates*, or *order parameters*, of ABMs [37,40].

Since tipping in stochastic ABMs is characterized by the existence of many metastable states and a multitude of transition pathways between them, we will apply Transition Path Theory (TPT) to the reduced model, and thereby gain a complete statistical understanding of the ensemble of transition paths between two chosen subsets A and B of the state space [43,70,73]. For studying noise-induced tipping between two metastabilities, one chooses A and B as metastable sets, i.e., two sets in which the system is trapped for a comparatively long time, but can eventually also escape from them. But TPT does not restrict A and B to be metastable, one can also study transitions between other relevant subsets of the state space, for instance given by viability constraints. TPT builds on the information that is contained in the *forward* and *backward committor functions*, i.e., in the hitting probability of B forward in time and of A backward in time. The advantage of studying tipping by TPT is that it allows to unravel the full range of transition pathways between sets A and B by computing the flux of transition paths, as well as other statistical properties of the transition paths, e.g., their distribution, rate and mean duration.

Recently, the forward committor has been singled out as the central object for quantifying the risk of future tipping [18,38]. Several papers study how one can solve for high-dimensional committors using neural networks [31,34,35,38]. Very much related to tipping is the concept of resilience, which in its simplest form is the system's ability of returning to the original state or region after a perturbation. Using similar objects as in TPT, namely escape probabilities and committors, this form of resilience of a system when in some or other attractor can be studied by analysing their stochastic basin of attraction [36,57,58].

TPT was originally developed for studying rare transitions in statistical mechanics, e.g., protein folding [46], and chemical reactions, but was later also applied for analysing transition events in the climate system [17] and marine debris dynamics [44]. Bifurcation diagrams of high-dimensional ABMs have already been studied [60,66], as well as the various bifurcation-induced transition pathways in a coupled social-ecological model [41], but to our knowledge noise-induced tipping in agent-based models has not been considered yet.

We will illustrate our approach on two paradigmatic models that exhibit tipping. The first model is based on Granovetter's threshold model [21] and describes the social activation of people for some collective action, such as rioting. Therein, when at least a certain fraction of an agents' neighbourhood is active in the collective action, the agent has a high chance of also becoming active. The second model considers agents that influence each other regarding their opinions and actual behavioural choices with respect to certain behavioural options, such as a more or less climate-friendly lifestyle or following certain epidemic countermeasures more or less stringently. The crucial feature of this model is that opinions and actual behaviours do not always have to agree. In particular, the model assumes that the more agents in the population hold the opinion that "one should do A" (e.g., wear a face mask), the more likely an agent can be convinced by her social peers to switch from choosing behaviour non-A to behaviour A for themselves. At the same time, the more agents in the population actually exhibit behaviour A, the more likely an agent can be convinced by her social peers to switch from the opinion "one should do A" to "one should not do A", since it may seem that the issue addressed by behaviour A is already sufficiently dealt with. This negative feedback loop then induces oscillatory dynamics. We will study both models on highly modular interaction networks, where the different blocks of agents (i.e., densely connected groups) influence each other. When the majority of agents in one block change their state, i.e., the block "tips", connected blocks are more likely to also tip. Thus tipping happens as a tipping cascade among connected blocks.

Our overall approach has the perspective of giving a quantitative analysis of noise-induced tipping without making any prior structural assumptions about the system. Instead of studying tipping points in bistable

systems or the stability of the attractors, Transition Path Theory offers a new perspective onto tipping by quantitatively characterising the dominant pathways along which tipping happens. This allows for a more detailed understanding of the tipping process especially for complicated systems, as well as for finding new ways of bypassing and preventing tipping. The introduced methodology could in the future be applied to more complex agent-based models.

In the following, we will first in Sect. 2 introduce two agent-based models that exhibit noise-induced tipping. In Sect. 3 we will show how we can find a reduced representation in terms of collective variables by using the Diffusion Maps algorithm. This allows us to finally in Sect. 4 analyse the tipping pathways using Transition Path Theory.

2 Two agent-based models exhibiting tipping

We start by introducing two agent-based models (ABMs) that exhibit tipping. The two presented models are part of the large class of models of opinion and behavioural change due to social dynamics [63]. The models in this class range from having discrete (e.g., the voter model [24]) to continuous states (e.g., the Deffuant-Weisbuch model [15]), as well as from having pair-wise interactions (also called simple contagion) to higher-order interactions (also named complex contagion) [5,9]. Often in these models one is interested in understanding the emergence of a stable macro-state of either opinion consensus or synchronous behaviour on the one hand, or opinion polarization or asynchronous behaviour on the other hand. In contrast to this, we are interested in the transitions between states of locally converged agents and have thus chosen models where the stochasticity enables tipping.

To be more specific, in our models agents are making binary behavioural decisions and change their binary opinions in reaction to the social influence of their network neighbours, potentially mediated by an additional macroscopic interaction (thus interactions are complex). Apart from their fixed position in the network, agents are identical. We will assume interaction networks consisting of several groups of nodes which are densely linked among themselves but with only few connections to the other groups. The densely linked agents in each block are nearly identical because they are connected to very similar sets of other agents, and thus behave rather similarly due to the local interaction rules. Thus both ABMs have many metastable states, where agents behave collectively in each of the blocks.

Many ABMs can be written as Markov chains or Markov jump processes, see [27] for some examples. The Markovianity assumption means that the next state of the system only depends on the current state and not the history.[1] The two models that we consider can be viewed as Markov chains $(\mathbf{X}_t)_{t\in\mathbb{Z}}$. They have a finite, but large state space and are irreducible, thus ergodic, as well as aperiodic. Due to these properties, both ABMs exhibit tipping, i.e., transitions between several metastable states. Later, we are interested in studying noise-induced tipping and tipping cascades between two diametrically opposed metastable regions of the dynamics. For the tipping analysis, we consider the models in stationarity.

For a comprehensive introduction to Markov chains we refer the reader to the book of Norris [47]. Note that we follow the convention to use uppercase letters X for random variables and lowercase letters x for their possible realizations.

2.1 A threshold model of social contagion or activation

We will introduce and discuss a very simple ABM of social contagion to describe phenomena such as the spreading of cultural fads, hypes or consumption behaviours, or the activation for some collective action such as rioting.

Let us consider a population of agents where each agent can be in one of two discrete states: being *inactive* or *active* in the collective action. The interaction topology between agents is given by a fixed network. A threshold-like influence is exerted by the social neighbours when an agent makes a binary decision: if more than a certain fraction of neighbours are in the opposite state to that of the agent, the agent will switch its state with a high probability. Thus each agent aligns its state with the state of the majority of its social neighbours. In addition, there is a small probability for the agent to switch its state without social influence, which can either be interpreted as a form of exploration or as representing otherwise unmodelled additional causes for switching one's state.

This ABM is ultimately based on Granovetter's threshold model, but Granovetter considered a fully mixed population [21]. More recently, several network-based versions of his original idea have been proposed [72,74], also containing different classes of agents such as stubborn agents that have a fixed state. Often, threshold distributions for the population are studied as well as deterministic interactions resulting in only one decision-making cascade through the population [21,72,74]. We instead consider the threshold to be constant for all agents and assign probabilities to the activity changes, thus our system can escape from the metastable regions.

Let us define our threshold model in more detail:

Interaction rules We consider a system of N interacting agents with social connections among them given by the

[1] This does not necessarily mean that agents have no memory or cannot be influenced by their past. By enlarging the state space formally to include a memory of past states, Markovian dynamics can be retained.

edges of a static network $\mathcal{G}$ of N vertices. The state of each agent i at the discrete time point t is denoted by $X_t^i \in \{0,1\}$ corresponding to being *inactive* or *active* in the collective action, respectively.

At each time $t = 0, 1, \ldots,$ each agent i in state $X_t^i = 0$ (resp. 1) will change their state to $X_{t+1}^i = 1$ (resp. 0)

- with probability p, if more than or exactly a fraction θ of neighbouring agents at time t are in the opposite state 1 (resp. 0),
- or with the exploration probability e, if less than a fraction θ of neighbours is in the opposite state,

where we assume $1 > p \gg e > 0$ such that social influence is stronger than exploration.

We can also view the system as a Markov chain $(\mathbf{X}_t)_{t\in\mathbb{Z}}$ on the state space $\mathbb{X} = \{0,1\}^N$, where we denote the *population state* at time t by $\mathbf{X}_t = (X_t^i)_{i=1}^N$. Since agents in every time step change their state synchronously and independently of each other, the transition matrix on $\mathbb{X}$ decomposes into the product of the "transition probabilities" for each individual agent

$$P(\mathbf{x},\mathbf{y}) := \mathbb{P}(\mathbf{X}_{t+1} = \mathbf{y} \mid \mathbf{X}_t = \mathbf{x}) = \prod_{i=1}^{N} \mathbb{P}(X_{t+1}^i = y^i \mid \mathbf{X}_t = \mathbf{x}). \quad (1)$$

The exploration probability ensures that agents are never stuck in a state. In every time step an agent has a positive probability to remain in the same state as well as to change the state, i.e., $\mathbb{P}(X_{t+1}^i = 0 \mid \mathbf{X}_t = \mathbf{x}) > 0$ and $\mathbb{P}(X_{t+1}^i = 1 \mid \mathbf{X}_t = \mathbf{x}) > 0$ respectively. Thus by (1) there is a positive probability to go from any population state to any other within one-time step, implying that the Markov chain is irreducible and also aperiodic.

Interaction network We assume that the interaction network $\mathcal{G}$ has two scales: it consists of *blocks*, sometimes also referred to as *communities* or *clusters*, in which the nodes are densely connected, whereas nodes of different blocks are sparsely connected. One popular approach to randomly generate such a network is by the stochastic block model. Each node i is uniquely assigned to a *block* $\mathcal{B}_k$, $k = 1, \ldots, K$. When node i belongs to $\mathcal{B}_k$, we also write $i \in \mathcal{B}_k$. After defining a symmetric matrix $W = (W_{kl})$ of size $K \times K$ that contains the edge wiring probabilities between a node of block $\mathcal{B}_k$ and one of block $\mathcal{B}_l$, we go through all pairs of nodes independently and with probability W_{kl} place an edge between them when they belong to blocks $\mathcal{B}_k$ and $\mathcal{B}_l$. The diagonal entries of W determine the edge wiring probabilities for agents from the same block. In the case of only one block, this is equivalent to the Erdős–Rényi random graph model.

Resulting dynamics If we consider a population where every agent is interacting with every other agent, i.e., they are interacting on a complete network[2], then in the interesting parameter regime the system switches between two metastable regions: (i) where a majority of agents is inactive and (ii) where most agents are active, see Fig. 1 for a realization with 50 agents. If we now study a population consisting of several (complete or mostly complete) blocks, with some connections between the blocks, then this structure is multiplied. Each block can switch between two such metastable regions, but depending on the number of connections between the blocks, all blocks are either synchronized, only weakly influencing each other, or behaving mostly independently. We set parameters to the case where tipping occurs and the blocks are weakly influencing each other to avoid a trivial behaviour and to focus on the most interesting dynamical regime.

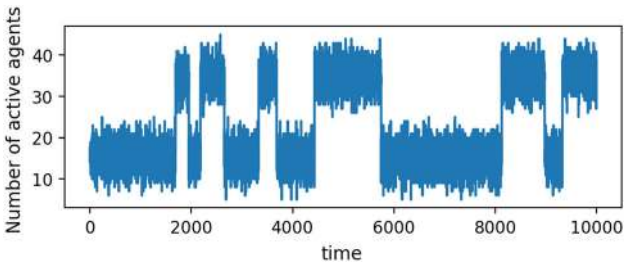

Fig. 1 Realization of the threshold model with 50 agents that are interacting on a complete network, i.e., each agent is influenced by the whole population. The dynamics switches between two metastable macrostates. The model parameters are $e = 0.3$, $p = 0.7$, $\theta = 0.5$

Example 1 The first example we will consider throughout the paper is a small population of just 10 agents that are evenly split into two blocks. We set the change probability as $p = 0.3$ and the exploration probability as $e = 0.03$. As long as $p \gg e$, the actual scale of the probabilities determines mostly how fast agents are changing their behaviour. The threshold was set to the most focal value of $\theta = 0.5$, meaning that agents are influenced by the majority behaviour in their neighbourhood. With a size of $|\mathbb{X}| = 2^{10}$, the state space is already nontrivial, but still small enough to be able to do direct computations of the state space. In Fig. 4a we show a realization on the small agent network that is shown in Fig. 4b. The realization indicates that the system remains in four metastable regions most of the time: where (i–ii) a majority in block 1 (resp. 2) but not in the other block is active, (iii) a majority in both blocks is inactive, and (iv) a majority in both blocks is active. It seems that those states (i–ii) where the two blocks show a differing majority activity are less metastable than those states (iii–iv) where agents in both blocks are conform. Moreover, the realization suggests that the tipping of one block induces the other block to also tip. By "tipping" we understand a transition from one metastable region to another, i.e., one

[2] We call a network complete if every node of the network is connected to every other node of the network.

block of agents drastically changes its state from the majority of agents active to majority inactive (or vice versa). Sometimes we might refer to the tipping of the whole population, i.e., when all blocks drastically change between the majority of agents being active and the majority being inactive. This happens via the individual blocks' successive tipping, i.e., via a *tipping cascade*.

Example 2 As a second example, we consider a large population structured into four blocks of different sizes. Block 1 contains 20 agents, all other blocks consist of 25 agents, see Fig. 9d for the network. The four blocks are circularly connected, and the network is generated by the stochastic block model where each agent has a wiring probability of 0.9 to agents in the same block and of 0.04 to agents from circularly neighboured blocks. We set $e = 0.23$, $p = 0.66$, $\theta = 0.5$. In this example, there are potentially 16 metastable regions, since agents in each block can be mostly active or not and there are four blocks. One can again assume that the tipping of one block induces neighbouring blocks to also tip, see Fig. 5a for a realization.

2.2 An oscillating, bivariate complex contagion model

In this second ABM, we are modeling the changes of binary opinions and separately of binary actual behavioural choices with respect to a certain behavioural option, such as a climate-friendly lifestyle or a certain preventive measure against an epidemic. For illustrative purposes, we use the context of climate-friendly lifestyles and the metaphor of "green" behaviour. We hence say that each agent has a *non-green* or *green* opinion, and also displays a *non-green* or *green* actual behaviour.

The model again considers a complex contagion process [9], where the social reinforcement from multiple agents at the same time is needed for an agent to change its state. But this time an agent's state has two components: opinion and actual behavioural, and the model also does not have a sharp threshold-like rule. Instead, the state change in opinion resp. actual behaviour of an agent is triggered upon interacting with two neighbours that both hold the opposite opinion resp. both display the opposite behaviour. Additionally, the actual probabilities with which these switches then occur also depend on the macroscopic state of the agent's block. In the model a switch in an agent's actual behaviour is made more likely by the respective opinion in the agent's block (e.g., the more agents have a *green* opinion, the more agents switch to a *green* behaviour), whereas a change of opinion is amplified if the block displays the opposite behaviour (e.g., the more agents display a *green* behaviour, the more agents will switch to a *non-green* opinion), the resulting dynamics leads to oscillations, i.e., is cyclic.

This model shows that opinions and actual behaviours do not always have to be aligned. There might be a time lag between holding a certain opinion and behaving accordingly. Additionally, the incentive to hold a certain opinion drops when many agents in the block are behaving in that way. It seems that there is no longer the need to hold the respective opinion since enough action is taken by other agents.

In more detail the model is formulated as follows:

Setting We consider a system of N agents, each agent i with a binary opinion $O_t^i \in \{0,1\}$ and a binary behaviour $B_t^i \in \{0,1\}$ at time t. For illustration we consider 0 as *non-green* and 1 as *green*. In each time step, each agent is interacting with two randomly drawn neighbours in a static social network $\mathcal{G}$. We again assume an interaction network with many communities, e.g, generated by the stochastic block model, and that each agent has at least two neighbours. Further, each agent i is influenced by the set of agents within the same block. For an agent $i \in \mathcal{B}_l$, we define the following *block* fractions:

$$\bar{O}_t^i := \frac{|\{j \in \mathcal{B}_l : O_t^j = 1\}|}{|\mathcal{B}_l|},$$

i.e., the fraction of agents with a *green* opinion in the same block as i, and

$$\bar{B}_t^i := \frac{|\{j \in \mathcal{B}_l : B_t^j = 1\}|}{|\mathcal{B}_l|},$$

the fraction of agents with a *green* behaviour in the same block. Note that these quantities, viewed as functions of the agents' index i, are constant on each block.

Below, the parameters $b, c \in [0,1]$ determine how strongly a *green* resp. *non-green* change in behaviour is influenced by the opinions in the block. Likewise, the parameters $f, g \in [0,1]$ determine how strongly a *green* resp. *non-green* change in opinion is influenced by the actual behaviour in the block. The general rate parameter $\tau \in (0,1)$ is for scaling the amount of change per time step.

Interaction rules At each discrete time point t, each agent i independently chooses two distinct neighbours j, k uniformly at random.

A behaviour change occurs:

- if $B_t^j = B_t^k = 1$, $B_t^i = 0$: agent i changes its behaviour to $B_{t+1}^i = 1$ with probability $\tau(b\,\bar{O}_t^i + (1-b))$,
- if $B_t^j = B_t^k = 0$, $B_t^i = 1$: agent i changes its behaviour to $B_{t+1}^i = 0$ with probability $\tau(c\,(1 - \bar{O}_t^i) + (1-c))$,
- or else, with a small exploration probability e, agent i changes its behaviour.

Thus an agent has a higher chance of changing its behaviour to *green* when interacting with two neighbours of *green* behaviour and the more likely the more

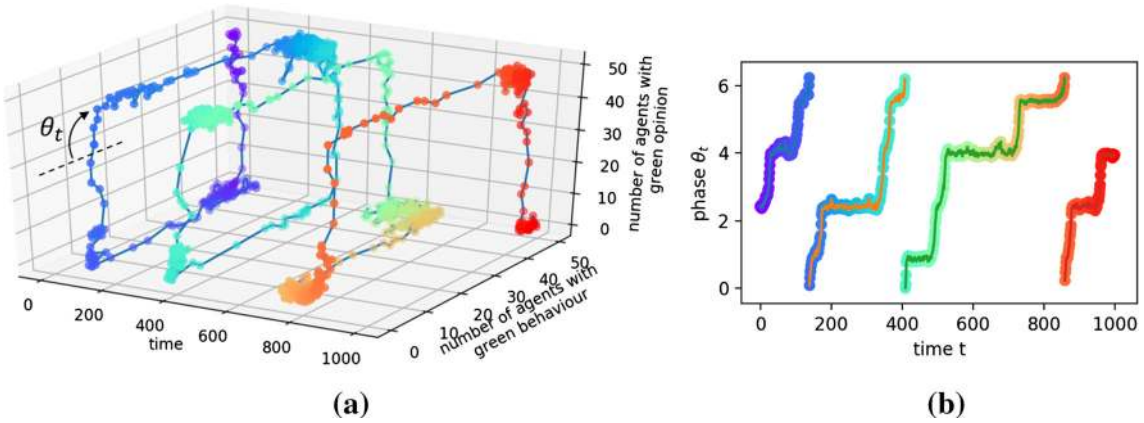

Fig. 2 Realization of the complex contagion dynamics for one complete network of 50 interacting agents. **a** The dynamics is strongly cyclic in the plane spanned by the two coordinates "number of agents with green opinion" and "number of agents with green behaviour". **b** Therefore, we can also visualize the dynamics by plotting the clockwise-angle in the coordinate plane, i.e., the phase θ_t. The model parameters are $b, c, f, g = 0.7$, $e = 0.02$, $\tau = 0.99$

agents in his block have a *green* opinion.[3] An agent is more likely to change its behaviour to *non-green*, when interacting with two neighbours of *non-green* behaviour and the more agents in his block show a *non-green* behaviour.

Conversely an opinion change happens:

- if $O_t^j = O_t^k = 1$, $O_t^i = 0$: with probability $\tau(f\,(1 - \bar{B}_t^i) + (1-f))$, agent i changes its opinion to $O_{t+1}^i = 1$,
- if $O_t^j = O_t^k = 0$, $O_t^i = 1$: with probability $\tau(g\,\bar{B}_t^i + (1-g))$, agent i changes its state $O_{t+1}^i = 0$,
- or else: with a small probability e, agent i changes its opinion.

This is now the other way around when an agent with a certain opinion (e.g., *green*) meets two neighbours of a different opinion (e.g., *non-green*) the change probability is higher the more agents in his block do not show this behaviour (i.e., the more show a *green* behaviour).

The exploration probability e should be small compared to τ. Since an agent first has to interact with two agents of a different state at the same time to have a higher chance for switching its state, it is hard for the dynamics to escape from a situation where agents in a block have converged. As a consequence, the dynamics are metastable. The exploration probability only offers a small chance for an agent to change its state.

The dynamics of the whole population can again be viewed as a Markov chain $(\mathbf{X}_t)_{t\in\mathbb{Z}}$ on the state space $\mathbb{X} = \{0,1\}^{2\times N}$, where we denote the population state at time t by $\mathbf{X}_t = (\mathbf{B}_t, \mathbf{O}_t) = (B_t^i, O_t^i)_{i=1}^N$. Requiring $0 < e, \tau < 1$ ensures that the Markov chain is irreducible and aperiodic.

[3] Note that if we disregard the rate τ, the first behaviour-change probability is a convex combination with factor b between the probabilities $\bar{O}_t^i$ ("fraction in block with green opinion") and 1 ("change with certainty to the behaviour of the two chosen neighbors").

Resulting dynamics If we consider a fully-connected population, in other words a complete network, and choose a large block influence strength $b, c, f, g = 0.7$, then the dynamics cycles in one direction through the four possible metastable regions where the large majority of agents share the same opinion and display the same behaviour (either the one aligned with the shared opinion or the opposite one), see Fig. 2a. Starting from a majority in the population with a non-green opinion and behaviour, first the majority changes their opinion to green, then after some time switches their behaviour also to green, followed by a change to a non-green opinion, and then also a non-green behaviour. Since the angle of rotation in the coordinate plane (also called *phase*) contains all information about the dynamics, we can essentially reduce the plot to 2 (b), where we show how the *phase* $\theta_t \in [0, 2\pi)$ varies in time. Whenever the phase remains approximately constant for some time, the system is in a metastable state.

For this model, we are at the end interested in the possible transitions from a majority of agents with *non-green* opinion and behaviour to a majority of agents with *green* opinion and behaviour, which is a succession of several transitions between metastable states, i.e., a tipping cascade.

Example 3 As an example throughout the paper, we consider a slightly larger population of 40 agents split into two blocks, see Fig. 6b for the network. In Fig. 6a we show for a short realization how the phase θ_t^j varies in time for each block $\mathcal{B}_j$. The block-wise phase is the angle of rotation at time t of the state in the coordinate plane of "number of agents with green opinion in block $\mathcal{B}_j$" vs "number of agents with green behaviour in block $\mathcal{B}_j$" as measured from the center point $(10, 10)$. We see that most of the time the phases of the two blocks are synchronized or mimicking each other, but they can also be completely out of synchrony. As model parameters we set $b = c = f = g = 0.7$, $e = 0.02$ and $\tau = 0.99$. The internal wiring probability in each block

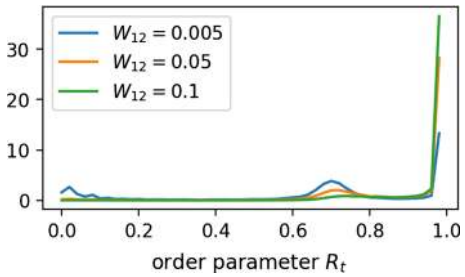

Fig. 3 Distribution over time of the order parameter R_t, Eq. (2), for the complex contagion model with different edge wiring probabilities W_{12} between the two blocks

is $W_{11} = W_{22} = 1$, thus both blocks are complete. An edge wiring probability of $W_{12} = W_{21} = 0.055$ between the two blocks ensures that the two blocks are mostly synchronized but still behave separately. The two blocks can also be viewed as two coupled oscillators where the coupling strength between oscillators is given by the edge wiring probability between the blocks.

In Fig. 3 we study the distribution of the order parameter

$$R_t = \Big| \sum_{j=1}^{2} \exp(\mathrm{i}\,\theta_t^j) \Big|, \tag{2}$$

in time, which is a basic measure of synchronization between coupled oscillators [1,49]. Here, i denotes the imaginary unit. In our case we compare the level of synchronization for different edge probabilities W_{12} between the two blocks. By placing every oscillator according to its phase θ_t^j on the unit circle, the order parameter measures the distance of the average of the positions on the unit circle from the origin. Thus when all oscillators are evenly spread out on the unit circle, R_t is close to 0, while when all oscillators are on the same spot, R_t is 1. The results in Fig. 3 confirm that for our chosen edge wiring probability the two blocks are synchronized most of the time.

3 Collective variables and reduced dynamics

One difficulty when analysing agent-based models lies in their high dimensionality. The size of the state space grows exponentially with the number of agents, and usually one is interested in studying a rather large population of agents. If the system state resides most of the time in the vicinity of some low-dimensional manifold, then we can search for *collective variables*, also called *reaction coordinates* or *order parameters*,

$$\xi : \mathbb{X} \to \mathbb{R}^d$$

that allow an approximate description of the actual system's dynamics in a "reduced" state space with much lower dimension d than that of the original agent state space $\mathbb{X}$. The reduced model approximately reproduces the emergent collective behaviour of the full ABM. The reduction allows us to better understand the structure of the dynamics and eases numerical computations.

Fortunately, the dynamics of many ABMs have a low intrinsic dimension. On the one hand, groups of people in many social situations are behaving rather collectively and are influenced by their peers, e.g., through copying and imitating their peers or the opposite, being repelled from their neighbours' behaviour. On the other hand, real-world social networks are often highly modular [19,67], i.e., contain many communities, as well as being scale-free, i.e., having a few nodes with very high degree, which additionally encourages coherent behaviour within sub-populations.

When agents are rather homogeneous or identical, one can usually guess suitable collective variables based on an intuition about the system's dominant feedbacks. Moreover, if the collective variables are a "simple" function of the agent state space, e.g., the number of agents that are in each of the different possible states, one can analytically derive mean-field approximations that take the form of coupled ODEs or SDEs [45,50] and whose continuous-time formulation simplifies an analytical treatment even further. In our two example ABMs, we have identical agents which are however heterogeneous due to their different positions in the network. This makes it more complicated to guess collective variables and also deteriorates the approximation quality of rather straightforward mean-field approximations. In many ABMs there are further forms of heterogeneity, such as varying interaction parameters.

We therefore seek an automated way of finding collective variables ξ, which should allow us to represent the dominant model behaviour, as well as all the dominant and important transition pathways between the metastable regions in state space. We have selected here TPT as a very promising approach for studying transition dynamics. To be able to apply TPT to the resulting reduced model, we must also make sure that the projections of these metastable regions onto the low-dimensional manifold spanned by the sought collective variables are well separated from each other.

A similar task is performed by nonlinear manifold learning approaches. Given a cloud of sampled data points, $\mathbb{D}$, they try to parameterize the nonlinear manifold from which the data has been sampled. In our case, we want to parameterize the manifold close to which a large share of the observed system states, $\mathbb{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$, lies once the system has become stationary. We will use the dominant coordinates produced by the Diffusion Maps algorithm [12,32] as collective variables (introduced in Sect. 3.1). Diffusion Maps have already been applied for finding collective variables of ABMs [37,40], and for the data-driven computation of dynamical quantities [65] such as committor functions. Diffusion Maps benefit from several properties that are needed for our application and that give them an advantage over other popular non-linear manifold learning methods. The algorithm is robust against noise

(compared to the Isometric Feature Mapping [53,64] and the locally-linear embedding [54]), is computationally not too expensive, and an out-of-sample extension exists, which can be employed for the interpolation of the non-linear coordinates on data points not contained in $\mathbb{D}$. Compared to the t-distributed stochastic neighbor embedding [69], the Diffusion Maps method is deterministic and always yields the same embedding.

As the ABM dynamics is described by the Markov chain's large microscopic transition matrix, it will be suitable to also describe the identified collective variables' time evolution by a similar but much smaller macroscopic transition matrix. To find this reduced transition matrix on the reduced state space $\xi[\mathbb{X}]$, we will discretise the projected space and perform a Monte-Carlo estimation of the transition matrix, see Sect. 3.3.

3.1 Diffusion maps

We want to apply Diffusion Maps [12,32] to a set of data points $\mathbb{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\} \subset \mathbb{X}$ with the goal of finding a low-dimensional projection $\xi : \mathbb{X} \to \mathbb{R}^d$ of the given sample. The dimension d should be small compared to the dimension of the original space $\mathbb{X}$. For this to work well, we assume that the data points approximately lie on a $d-$dimensional manifold $\mathcal{M}$ embedded in the high-dimensional space $\mathbb{X}$.

The general idea of Diffusion Maps is to define a random walk on the data points $\mathbb{D}$, where the transition probability between similar or near points is high and between far points is close to zero. The random walk traverses the manifold and only follows its intrinsic structure, since the distances between near points in the original space are a good approximation to the local distances on the manifold. The dominant eigenvectors of the resulting transition matrix scaled by the corresponding eigenvalues can then be used as a nonlinear projection.

The transition matrix on the data points $\mathbb{D}$ is constructed as follows:

1. Choose a kernel $k^\epsilon(\mathbf{x}, \mathbf{y}) = h\left(\frac{d(\mathbf{x},\mathbf{y})^2}{\epsilon}\right)$ that describes the similarity of two data points, for example the popular Gaussian kernel given by $h(z) = \exp(-z)$. Moreover one has to set the scale parameter $\epsilon > 0$, e.g., by using the heuristic from [6,32], and a distance function $d(\cdot,\cdot)$ that is suitable for the data set, e.g., the Euclidean or Mahalanobis metric.
2. Letting $q^\epsilon(\mathbf{x}_i) = \sum_{m=1}^M k^\epsilon(\mathbf{x}_i, \mathbf{x}_m)$, we form the new anisotropic kernel

$$\tilde{k}^\epsilon(\mathbf{x}_i, \mathbf{x}_j) = \frac{k^\epsilon(\mathbf{x}_i, \mathbf{x}_j)}{q^\epsilon(\mathbf{x}_i)\, q^\epsilon(\mathbf{x}_j)},$$

   which has some desirable properties compared to k^ϵ, for more details see [12].
3. Applying row-normalization by $d^\epsilon(\mathbf{x}_i) = \sum_{m=1}^M \tilde{k}^\epsilon(\mathbf{x}_i, \mathbf{x}_m)$, we arrive at the transition matrix

$$P^\epsilon(\mathbf{x}_i, \mathbf{x}_j) = \frac{\tilde{k}^\epsilon(\mathbf{x}_i, \mathbf{x}_j)}{d^\epsilon(\mathbf{x}_i)}.$$

The matrix P^ϵ can be interpreted as the normalized Laplacian of a weighted undirected graph whose weights correspond to the anisotropic kernel $\tilde{k}^\epsilon$. As such it is reversible with respect to the stationary distribution $\pi(\mathbf{x}_i) = \frac{d^\epsilon(\mathbf{x}_i)}{\sum_j d^\epsilon(\mathbf{x}_j)}$.

The right eigenpairs (λ_j, ψ_j), $j = 0, \ldots, M-1$ of P^ϵ contain information about the geometric structure of $\mathbb{D}$ at different scales and are real-valued due to P^ϵ being reversible. We order the eigenpairs by decreasing magnitude of their eigenvalues. Then the leading eigenvectors, i.e., with the largest eigenvalues in magnitude, scaled by their corresponding eigenvalue, are a good projection of the large-scale structures in the data

$$\xi(\mathbf{x}_i) = (\xi_{1,i}, ..., \xi_{d,i}) = (\lambda_1 (\psi_1)_i, ..., \lambda_d (\psi_d)_i) \in \mathbb{R}^d,$$

where $(\psi_j)_i$ is the ith component of the jth eigenvector. Since the eigenvector corresponding to the largest eigenvalue is just the 1-vector and contains no information, we exclude it from the projection. As d we usually choose the number of remaining eigenvalues above the spectral gap. The Euclidean distances in these coordinates approximately correspond to the local diffusion distances on the manifold.

The computational cost of computing pair-wise distances and the eigenvectors of P^ϵ becomes very expensive if not impossible for very large data sets. To circumvent that, one can sub-sample the data set, compute the diffusion matrix and eigenpairs only for the sub-sample and interpolate the computed eigenvectors at the remaining data points with the help of the out-of-sample extension [13]. We refer the reader to [32] for an explanation of the extension.

3.2 Collective variables of the two ABMs

Next we show the results of using the dominant Diffusion Map coordinates as collective variables for our two ABMs. We used the Diffusion Maps implementation from [61] and applied the algorithm to a sample of 20,000 population states $\mathbb{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{20,000}\}$. As a kernel we used the Gaussian kernel and computed the distance $d(\mathbf{x}_i, \mathbf{x}_j)$ between two data points via the Hamming distance, which measures the distance between two binary strings as the number of entries where they differ and is therefore suitable for binary population vectors. Further, we estimated an appropriate scale parameter ϵ using the heuristic from [6].

Threshold model Since we set up the model such that agents in each block are nearly indistinguishable, the obvious choice for collective variables for this system is just the number of active agents in each block (or equiv-

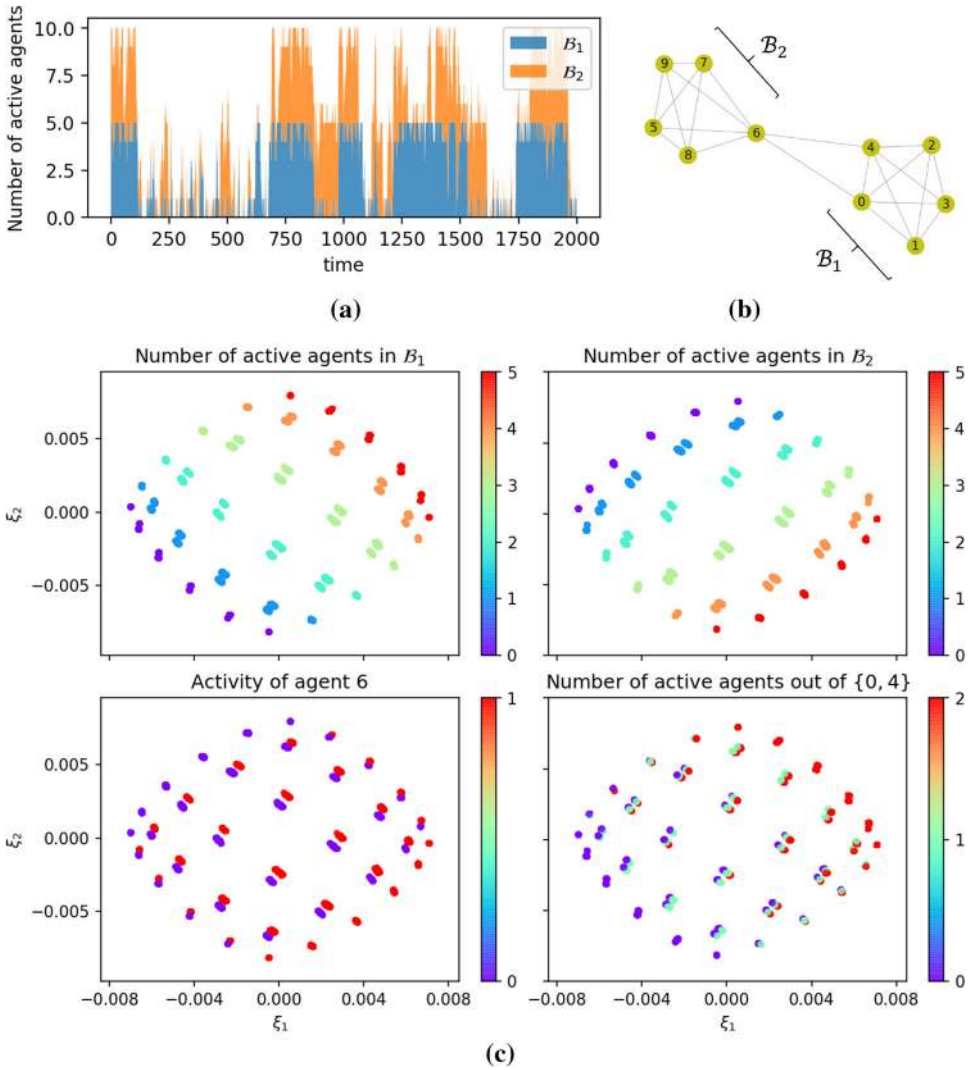

Fig. 4 Threshold model with two blocks of 5 agents each as in *Example 1*: **a** the realization is shown using a stackplot, i.e., the number of active agents in $\mathcal{B}_2$ is plotted vertically on top of the number of active agents in $\mathcal{B}_1$. Several tipping events are shown. **b** Modular agent network of two blocks. **c** Projection of population states into the dominant two Diffusion Maps coordinates, the Diffusion Maps scale parameter turned out to be $\epsilon = 0.25$. To better understand the projection, the data points are colored according to the number of active agents in each block and the activity of agents $0, 4$ and 6

alently, inactive agents). So let us see how the Diffusion Maps algorithm projects the data.

Example 1 continued: The projection into the dominant two coordinates can be found in Fig. 4. The sample of 20,000 population states are embedded into a square. The coloring of the data points indicates that the two orthogonal directions encode the number of active agents in each block. Moreover, note that the first Diffusion Map coordinate encodes the total number of active agents in the population and the second refines this by splitting them into two blocks. The Diffusion Map coordinates are refining the structure of the manifold with each additional coordinate and are ordered by the scales they encode. Looking more closely, we can see that the projected groups of points (corresponding to a certain number of active agents in each block) consist of some substructures on a smaller scale. These substructures encode whether agent 6 is active or not, and how many of agent 0 and 4 are active (see Fig. 4). Higher-order Diffusion Maps coordinates, in this case the coordinate ξ_4, also decode the information about the activity of agents $0, 4$ and 6 (not shown in the figure). We will later investigate the importance of agents 0, 4 and 6 with respect to the dynamics.

Judging from the location of the spectral gap of the Diffusion Maps spectrum, the intrinsic dimension of the dynamics seems to coincide with the number of blocks, i.e., $d = 2$.

Example 2 continued: The projection onto the two most dominant coordinates out of the four dominant ones can be found in Fig. 5. Though it is not so easy to see from just the dominant two coordinates, the data set

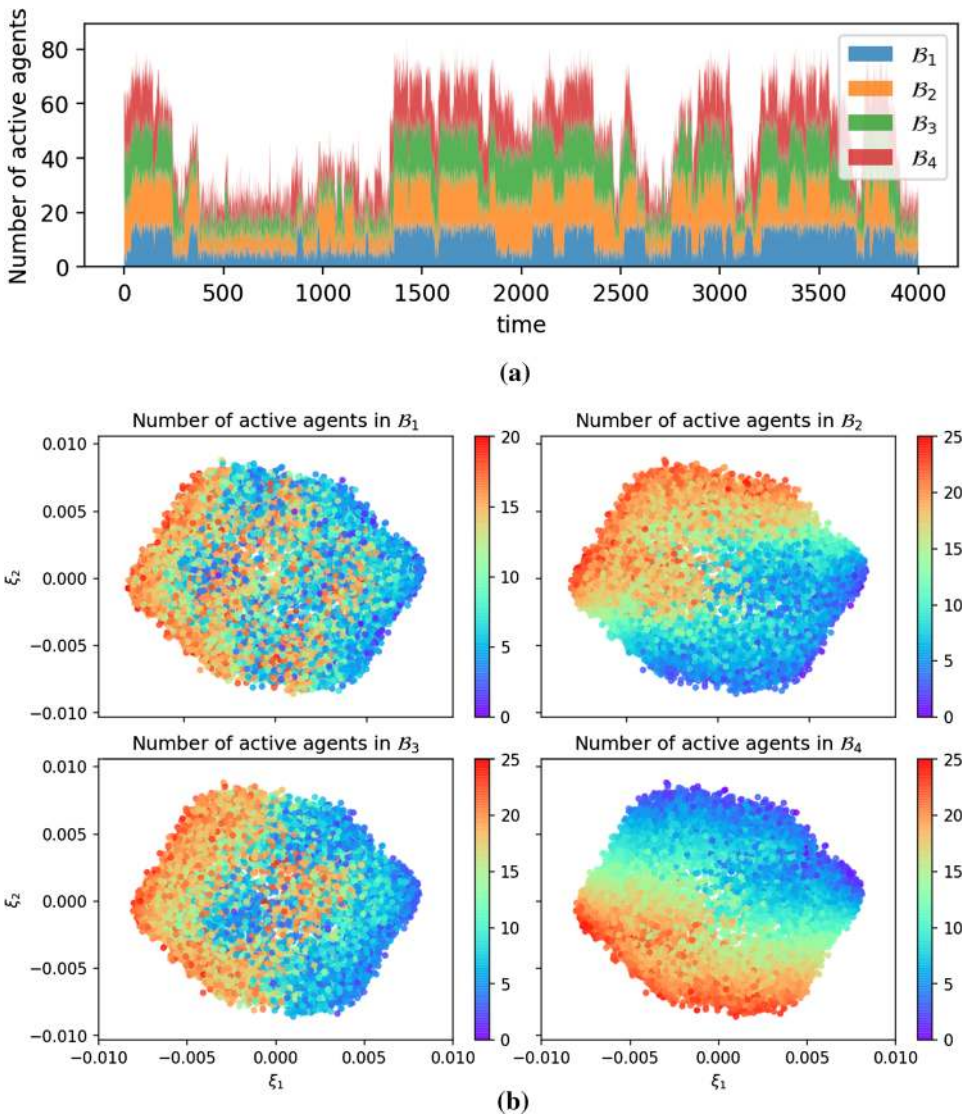

Fig. 5 Threshold model dynamics with four incomplete blocks as in *Example 2:* **a** the realization is shown as a stackplot. Several small tipping events (where agents in just one block switch their state) as well as tipping cascades (where nearly all agents change their activity) are apparent. **b** The diffusion maps projection into the first two coordinates is colored according to the number of active agents in each block which suggests the tesseract structure. The Diffusion Maps scale parameter came out as $\epsilon = 0.15$

of 20,000 samples from a long realization are embedded into a four-dimensional hypercube, a *tesseract*, whose corners correspond to the states where the majority of agents in a certain set of blocks are active and the others not. The edges of the hypercube are much less visible but also present. They are not visited that frequently, since they correspond to the rare transitions between metastable regions. For our computations later we will use all four Diffusion Map coordinates.

Remark Note that the Diffusion Maps algorithm approximately preserves the local distances on the manifold on which the dynamics takes place. In some situations, e.g., for visualization purposes, it might be of interest to find even lower-dimensional embeddings that do not necessarily preserve the local distances but are still non-overlapping. For instance, the net of the 3-D cube can be embedded into 2-D without overlapping edges by a planar graph projection.

Complex contagion model For this model one would guess that the number of agents with green opinion in each block and the number of agents with green behaviour in each block will constitute good collective variables. Or, to reduce it further since the blocks behave like coupled oscillators, one could try using only these oscillators' phase angles in the plane spanned by the number of agents with green behaviour and opinion in each block.

Example 3 continued: In Fig. 6 we show the Diffusion Maps projection into the dominant three coordinates, though four coordinates are needed to describe the dominant dynamics as indicated by the number of dominant eigenvalues of P^ϵ. Still, the three dominant coordinates in the figure already visually indicate that the data are projected onto a 4-D hypercube.

Diffusion Maps is not designed to embed a circle into a 1-D torus, only into a 2-D Euclidean space. Similarly, Diffusion Maps cannot embed the tesseract onto a 2-D torus. Here we will try to further post-process the embedding of $\mathbb{D}$ into $\mathbb{R}^4$ and project the net of the tesseract onto a 2-D torus. The tesseract can be projected onto the 2-D torus without crossing edges. We choose two two-dimensional planes in $\mathbb{R}^4$ that are orthogonally meeting in the center of the projected

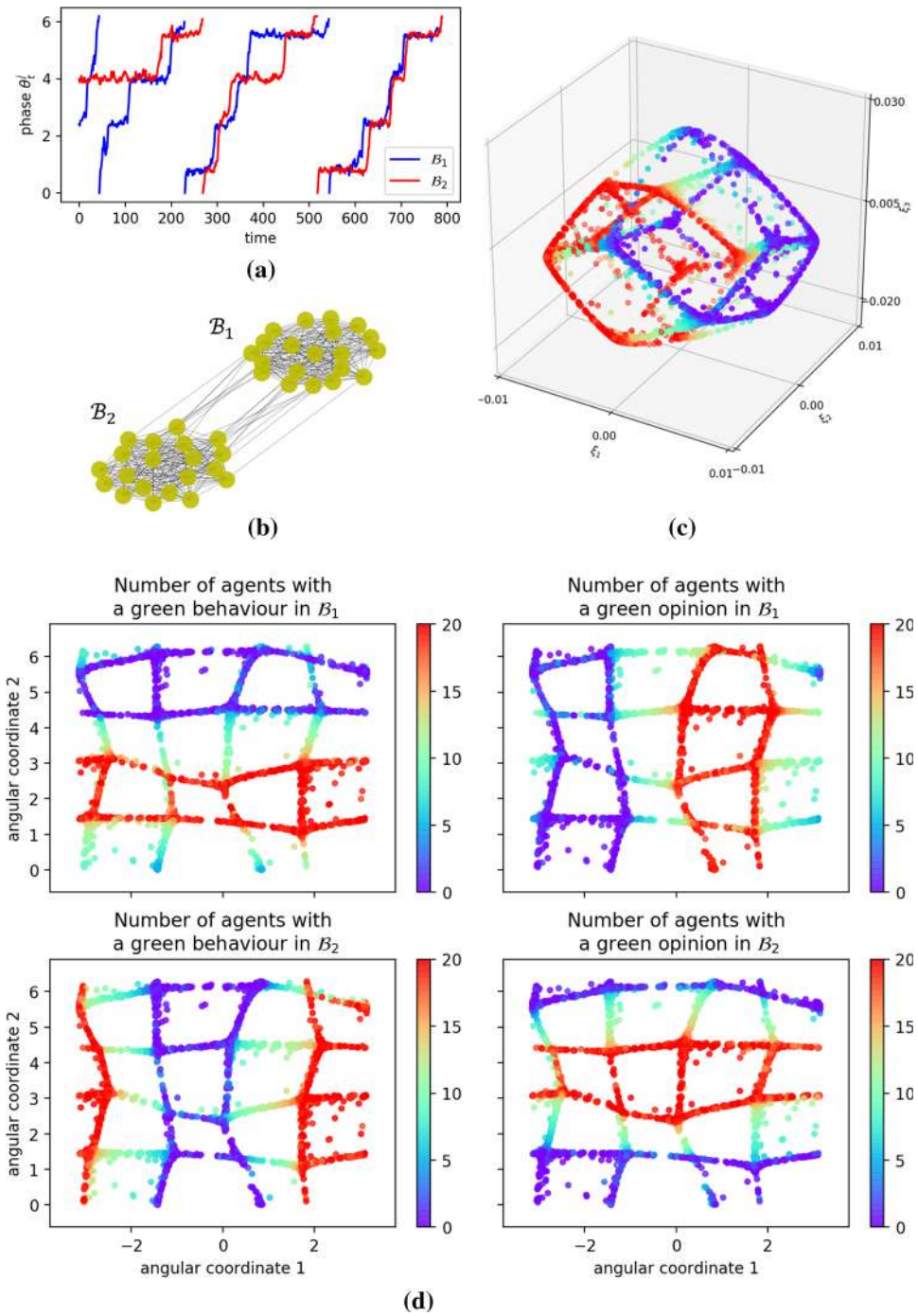

Fig. 6 Dynamics of the complex contagion model with two blocks as in *Example 3*: **a** for both blocks we plot how the phase θ_t^j varies in time. Most of the time the two blocks have a synchronized phase, when one block changes its state, the other block follows with some time lag. But it is also possible (see the beginning of the realization) that the two blocks are completely out of phase. **b** Network. **c** Projection of the data set into the first three Diffusion Maps coordinates with scale parameter $\epsilon = 0.1$, the data is projected into the skeleton of a tesseract. **d** Post-processing of the $4D$ Diffusion Maps projection (first three dimensions are shown in **c**) by visualizing the data points on a torus (this means that the opposite sides of the plot are identified with each other) and thereby unfolding the tesseract net. The data points are colored according to the number of agents of a certain opinion and behaviour in each block

tesseract and such that when measuring the angles in these planes, we can untangle the net of the tesseract without edges crossing each other. See Fig. 6 for the resulting net of the tesseract on the 2-D torus. All the computations will still be done using the four Diffusion Maps coordinates, the projection onto the 2-D torus is only for visualization purposes.

Even though the projection indicates that the dynamics essentially takes place on a tesseract, we yet cannot infer how the dynamics moves along the edges of the tesseract. We know the model is strongly non-reversible, so on the edges there will be a dominant direction of the probability flux.

3.3 Estimating the dynamics on the projected space

To study the reduced dynamics and apply Transition Path Theory to a Markov chain, we need its transition matrix. We will explain how one can estimate a low-dimensional transition matrix that describes the dynamics on the projected space $\xi[\mathbb{X}]$ from simulation data of the ABM.

We assume that we have sampled i.i.d. pairs of consecutive states $(\mathbf{x}^k, \mathbf{y}^k)_{k=1}^K$ of the Markov chain, in our case of the ABM. This means that $\mathbf{x}^k$ is sampled from the stationary distribution π and $\mathbf{y}^k$ is sampled from the conditional transition probabilities $P(\mathbf{x}^k, \cdot)$. Their projection into collective variables is given by $(\xi(\mathbf{x}^k), \xi(\mathbf{y}^k))_{k=1}^K$.[4] After partitioning the projected state space into M Voronoi cells $\{V_1, \ldots, V_M\}$, e.g., by using the K-Means clustering algorithm for finding the cell centers, we can estimate a transition matrix $P^\xi(m,n) = \mathbb{P}(\xi(\mathbf{X}_{t+1}) \in V_n \mid \xi(\mathbf{X}_t) \in V_m)$ on the state space $\mathbb{S} = \{1, \ldots, M\}$ identified with the Voronoi cells. Compared to regular box discretizations, Voronoi cells allow a discretization that is better fitted to the distribution and geometry of the trajectory data. We estimate P^ξ using *Ulam's method* by counting the proportion of transitions that went from V_m to V_n within one time step [42,56,68]:

$$\hat{P}^\xi(m,n) = K^{-1} \sum_{k=1}^{K} \mathbb{1}_{V_m}(\mathbf{x}^k)\, \mathbb{1}_{V_n}(\mathbf{y}^k).$$

Instead of using many one-step trajectories, one can also use one long ergodic trajectory and count the one-step transitions therein. But for systems with many metastable regions, the ergodic trajectory needs to be very long to correctly sample the stationary density and to attain a good estimate of the transition probabilities.

4 Studying tipping

When we are interested in the transitions from one subset $A \subset \mathbb{S}$ to another subset $B \subset \mathbb{S}$, we can study the ensemble of trajectory pieces that start in A, end in B, and in between only pass states in $C := \mathbb{S} \backslash (A \cup B)$, the so-called *reactive trajectories*. TPT is a framework to get statistical information, e.g., the rate, density, flux and mean duration, about the reactive trajectories [43,70,73]. If A and B are chosen as two metastable sets, then TPT studies the noise-induced tipping between metastable regions. But A and B can be any meaningful sets between which one wants to study the transition dynamics. TPT becomes particularly useful when tipping is very uncertain and there is a multitude of pathways linking A to B.

The forward and backward committors contain all the essential information about the possible future and past transitioning behaviour and generate the various statistics of the ensemble of reactive trajectories. The *forward committor* gives the probability to first reach B not A when starting in $x \in \mathbb{S}$

$$q^+(x) := \mathbb{P}(\tau_B^+(t) < \tau_A^+(t) \mid X_t = x),$$

where the random variable $\tau_S^+(t) := \inf\{s \geq t : X_s \in S\}$ is the first hitting time of the set $S \subset \mathbb{S}$ with the convention $\inf \emptyset := \infty$. The *backward committor* gives the probability to have last visited A not B when currently in x,

$$q^-(x) := \mathbb{P}(\tau_A^-(t) > \tau_B^-(t) \mid X_t = x),$$

where we denoted the last exit time of the set $S \subset \mathbb{S}$ by $\tau_S^-(t) := \sup\{s \leq t : X_s \in S\}$, $\sup \emptyset := -\infty$. Due to the assumption of a stationary process, the committors are time-independent.

Since we are in a stochastic setting, where tipping is usually not certain, we cannot define an interesting tipping point as a point of no return. Instead, the tipping points or edge states can be identified with the points where tipping to B is as likely as going back to A, i.e., those states with $q^+(x)$ close to $1/2$. Recently it has been argued that the forward committor is the relevant object to quantify the risk of tipping to a certain state in the future [18,38].

In this section, we study ABMs described by a discrete Markov chain $(X_t)_{t\in\mathbb{Z}}$ on the reduced state space $\mathbb{S} = \{1, \ldots, M\}$ that are stationary, irreducible and aperiodic. The ABM dynamics on $\mathbb{S}$ is described by the transition matrix P^ξ. In the following we will for simplicity write P.

In the following, we will first introduce TPT before applying it to our two ABMs and studying their tipping behaviour in depth.

4.1 Transition path theory

We are interested in characterizing the transitions from one subset of the state space $A \subset \mathbb{S}$ to another $B \subset \mathbb{S}$ (both non-empty and disjoint) in an irreducible and aperiodic Markov chain $(X_t)_{t\in\mathbb{Z}}$ on a finite state space $\mathbb{S}$. The time-homogeneous transition probabilities between states x and y are given by $P(x,y) = \mathbb{P}(X_{t+1} = y \mid X_t = x)$. Since we assume the process to be stationary, the distribution for all $t \in \mathbb{Z}$ is given by the *stationary probability distribution* π, the unique probability-vector solution to $\pi^\top P = \pi^\top$.

TPT provides statistics about these transitions linking A to B in a stationary Markov chain by making use of the information contained in the *forward committor* $q^+(x)$, the probability to first hit B rather than A when starting in $x \in \mathbb{S}$, and the *backward committor* $q^-(x)$, the probability to have last come from A not B when in x. The *forward committor function* q^+ uniquely solves

$$\begin{cases} q^+(x) = \sum\limits_{y\in\mathbb{S}} P(x,y)\, q^+(y) & x \in C \\ q^+(x) = 0 & x \in A \\ q^+(x) = 1 & x \in B, \end{cases} \tag{3}$$

while the *backward committor function* q^- is the unique solution to the linear system

[4] In our examples, we will apply Diffusion Maps to a sub-sample of the states $(\mathbf{x}^k, \mathbf{y}^k)_{k=1}^K$ and interpolate ξ at the remaining data points with the Diffusion Maps out-of-sample extension.

$$\begin{cases} q^-(x) = \sum_{y\in\mathbb{S}} P^-(x,y)\, q^-(y) & x \in C \\ q^-(x) = 0 & x \in B \\ q^-(x) = 1 & x \in A \end{cases} \tag{4}$$

where $P^-(x,y) = \mathbb{P}(X_{t-1} = y \mid X_t = x) = \frac{\pi(y)}{\pi(x)} P(y,x)$ are the backward-in-time transition probabilities.

Next, we will define some statistics of the ensemble of reactive trajectories. As a *reactive trajectory* we consider trajectory snippets $(x_t, x_{t+1}, \ldots, x_{t+T})$ that start in $x_t \in A$, end in $x_{t+T} \in B$ and in-between only pass through $x_{t+1}, \ldots, x_{t+T-1} \in C$. The excursion $(x_{t+1}, \ldots, x_{t+T-1})$ through the transition region C is called an *inner reactive trajectory*. When on a reactive trajectory it holds that both $\tau_A^-(t) > \tau_B^-(t)$ and $\tau_B^+(t+1) < \tau_A^+(t+1)$, while on an inner reactive trajectory we have $\tau_A^-(t) > \tau_B^-(t)$ and $\tau_B^+(t) < \tau_A^+(t)$. The ensemble of reactive trajectories is the collection of all reactive trajectory pieces that can be pruned out from the ensemble of stationary trajectories.

The *distribution* of inner reactive trajectories is defined as

$$\begin{aligned} \mu^{AB}(x) &= \mathbb{P}(X_t = x, \tau_A^-(t) > \tau_B^-(t), \tau_B^+(t) < \tau_A^+(t)) \\ &= q^-(x)\, q^+(x)\, \pi(x) \end{aligned}$$

and indicates where inner reactive trajectories spend most of their time, and thus also the bottlenecks during transitions. The function μ^{AB} is not normalized, but can be normalized by dividing by

$$Z^{AB} = \mathbb{P}(\tau_A^-(t) > \tau_B^-(t), \tau_B^+(t) < \tau_A^+(t)),$$

the probability to be on an inner reactive trajectory at time t.

The *current* or *flux* of reactive trajectories f^{AB} denotes the probability of a reactive trajectory to visit x and y consecutively:

$$\begin{aligned} f^{AB}(x,y) &= \mathbb{P}(X_t = x, X_{t+1} = y, \\ &\qquad \tau_A^-(t) > \tau_B^-(t), \tau_B^+(t+1) < \tau_A^+(t+1)) \\ &= q^-(x)\, \pi(x)\, P(x,y)\, q^+(y). \end{aligned}$$

Again, this function is not normalized w.r.t. the pair (x,y) and gives only the share of probability current accounting for reactive trajectories from A to B. While the usual probability current $\mathbb{P}(X_t = x, X_{t+1} = y)$ sums to 1, the reactive current sums to

$$\mathbb{P}(\tau_A^-(t) > \tau_B^-(t), \tau_B^+(t+1) < \tau_A^+(t+1)) =: H^{AB},$$

which is the probability to be on a reactive trajectory.

The reactive current into a state $x \in C$ equals the reactive current out of that state. A and B act as a source and sink of the reactive current, and the total reactive current out of A equals the reactive current into B:

$$\sum_{x\in A, y\in\mathbb{S}} f^{AB}(x,y) = \sum_{x\in\mathbb{S}, y\in B} f^{AB}(x,y).$$

We denote this quantity by k^{AB}, it specifies the *rate* of reactive trajectories, since it estimates the number of reactive trajectories that are started in A per time step, or equivalently, that end in B per time step. Further, by dividing the probability to be reactive by the transition rate, we obtain the mean duration of an inner reactive trajectory,

$$t^{AB} = \frac{Z^{AB}}{k^{AB}}.$$

For us, the reactive current is the most important quantity, since it reveals the multitude of transition pathways from A to B and their respective weight. Often one is interested in cycle-erased reactive trajectories from A to B, which are free from uninformative cycles and only contain the progressive parts from A to B. In the case of reversible dynamics, the effective current, which gives the net amount of reactive current from state x to y:

$$f^+(x,y) := \max\{f^{AB}(x,y) - f^{AB}(y,x), 0\}$$

reduces to $\pi(x)P(x,y)(q^+(y) - q^+(x))$ for $q^+(y) > q^+(x)$ and 0 else, and is therefore free of cycles.[5] For non-reversible processes, such as the ones we study, the effective current is however not guaranteed to be cycle-free and we have to take a different approach. Moreover, it might be interesting to study the cycles of reactive trajectories separately.

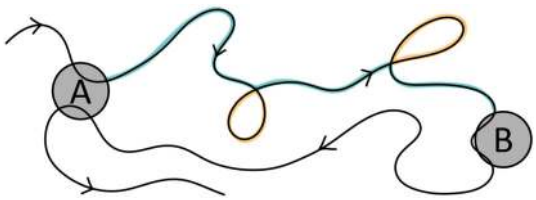

Fig. 7 Splitting of a reactive trajectory into one productive path from A to B (in blue) and several unproductive cycles (yellow)

Flux decomposition into productive and unproductive parts In [4] it is proposed to decompose the reactive flux $f^{AB} = f^P + f^U$ into the flux f^P coming from productive, cycle-free paths linking A to B and the flux f^U from unproductive cycles. In Fig. 7 we sketch the productive piece of a reactive trajectory linking A with B as well its two unproductive cycles.

Denote by Γ^P the set of non-intersecting paths $\gamma = (x_1, \ldots, x_s)$ that start in $x_1 \in A$, end in $x_s \in B$, and go

[5] If this were not the case, then the committor would have to strictly increase along a cycle for $f^+(x,y)$ to be positive, but this is impossible.

through the transition region $x_2, .., x_{s-1} \in C$. By non-intersecting we mean that all of the traversed states x_r are pairwise different, thus ensuring that the path is free of cycles. Further, all the visited edges (x_r, x_{r+1}) along the path need to have a positive transition probability $P(x_r, x_{r+1}) > 0$. By Γ^U we denote the set of non-intersecting paths $\gamma = (x_1, \dots, x_s)$ through the transition region $x_r \in C$ that are closed. Since the path is closed, it additionally contains the edge (x_s, x_1). Non-intersecting, closed paths are also called *cycles*. Note that self-cycles $\gamma = (x)$, i.e. paths that go from x to x, are also considered as cycles.

Now we are equipped to decompose the reactive current into the current from cycle-free productive paths Γ^P and the current from unproductive cycles Γ^U [4]

$$f^{AB}(x,y) = \underbrace{\sum_{\gamma \in \Gamma^P} w(\gamma)\, C^\gamma(x,y)}_{f^P} + \underbrace{\sum_{\gamma \in \Gamma^U} w(\gamma)\, C^\gamma(x,y)}_{f^U},$$

where C^γ is the incidence function of the path γ

$$C^\gamma(x,y) = \begin{cases} 1, & \text{if } \gamma = (\dots, x, y, \dots) \\ 0, & \text{else} \end{cases}$$

and $w(\gamma)$ encodes the path weight. $w(\gamma)$ can be understood as an average along an infinitely long ergodic trajectory $(x_t)_{t\in\mathbb{N}}$ as follows

$$w(\gamma) = \lim_{T\to\infty} \frac{N_T^\gamma}{T}, \tag{5}$$

where N_T^γ counts the number of times that $(x_t)_{t=1,\dots,T}$ passes through γ while reactive. The edges of γ have to be passed in the right order but excursions to one or more other cycles in between are allowed.[6]

This decomposition into a productive and unproductive flux allows an interpretation of how much reactive trajectories are passing the different paths and cycles. The easiest way to numerically estimate this decomposition is as follows:

1. Sample a long trajectory $(x_t)_{t=1,\dots,T}$ that contains sufficiently many transitions from A to B. Since we only need the reactive trajectory pieces for the computation of (5) but correctly weighted by H^{AB} compared to the non-reactive pieces, one can also use a transition matrix that only samples the reactive pieces of the trajectory correctly and maps all the non-reactive pieces to a single state [8,71].
2. Estimate $w(\gamma)$ by averaging along this sample trajectory [3]. First prune out all the reactive pieces. Then for each reactive snippet iteratively cut out all the cycles by going through the trajectory until for the first time a state is revisited, i.e., until we find r such that $x_r = x_m$, $m < r$. Take out the cycle $(x_m, \dots x_{r-1}) = \gamma$ and increment N_T^γ by 1. Repeat until from the reactive snippet only a cycle-free transition path γ is left, increment N_T^γ accordingly. Then move to the next reactive trajectory piece.

Flux aggregation In order to assess the flux through certain subsets of the state space, e.g., through different channels or other regions of the state space, we can aggregate states together and compute the reaction current between aggregate region. First we partition the state space into groups of states $L_s \subset \mathbb{S}$, in such a way that the disjoint union of elements in $\{L_1, \dots, L_S\}$ is the whole state space $\mathbb{S}$ and that the boundaries of A and B are preserved. Then we can compute the *reactive macro-current* F^{AB} between the group of states L_r and L_s as follows [46]:

$$F^{AB}(r,s) = \sum_{x\in L_r, y\in L_s} f^{AB}(x,y).$$

These macro-currents fulfill the same properties as the micro-currents, namely, the total flux out of A equals the total flux into B, moreover the flux into a partition element L_r equals the flux out of that partition element L_r. We can also compute the effective macro-current F^+ between partition elements from the reactive macro-current.

4.2 Tipping analysis of the ABMs

In this section, we apply TPT to the two reduced ABMs to understand the possible tipping pathways.

When the two presented ABM dynamics are considered for modular agent networks, they have many metastabilities of different quality and strength. A multitude of transition paths exists that go from some region of interest A, via several stopovers in metastable regions, to some other region B. These transition pathways from A to B are forming a transition network. We can understand the tipping dynamics also in terms of tipping cascades among connected blocks: when one block tips, it influences the probability of other connected blocks to also tip.

Threshold model For the threshold model we are interested in studying how the activity in collective behaviour spreads through the population. Therefore, we will set A as all the states where a small proportion of agents is active and B as all the states where the majority of agents in the population is active. Note that we have to be able to express A and B via the collective variables. We used the TPT code from [23].

Example 1 continued: We discretized the space spanned by ξ into 36 Voronoi cells, and estimated the transition

[6] The original derivation in [4] proceeds slightly differently. They modify the reactive current to also include current from B to A, and then apply the *stochastic cycle decomposition* [28,29] for conserved currents to uniquely decompose the modified current into cycles solely in C and cycles that contain an edge from B to A.

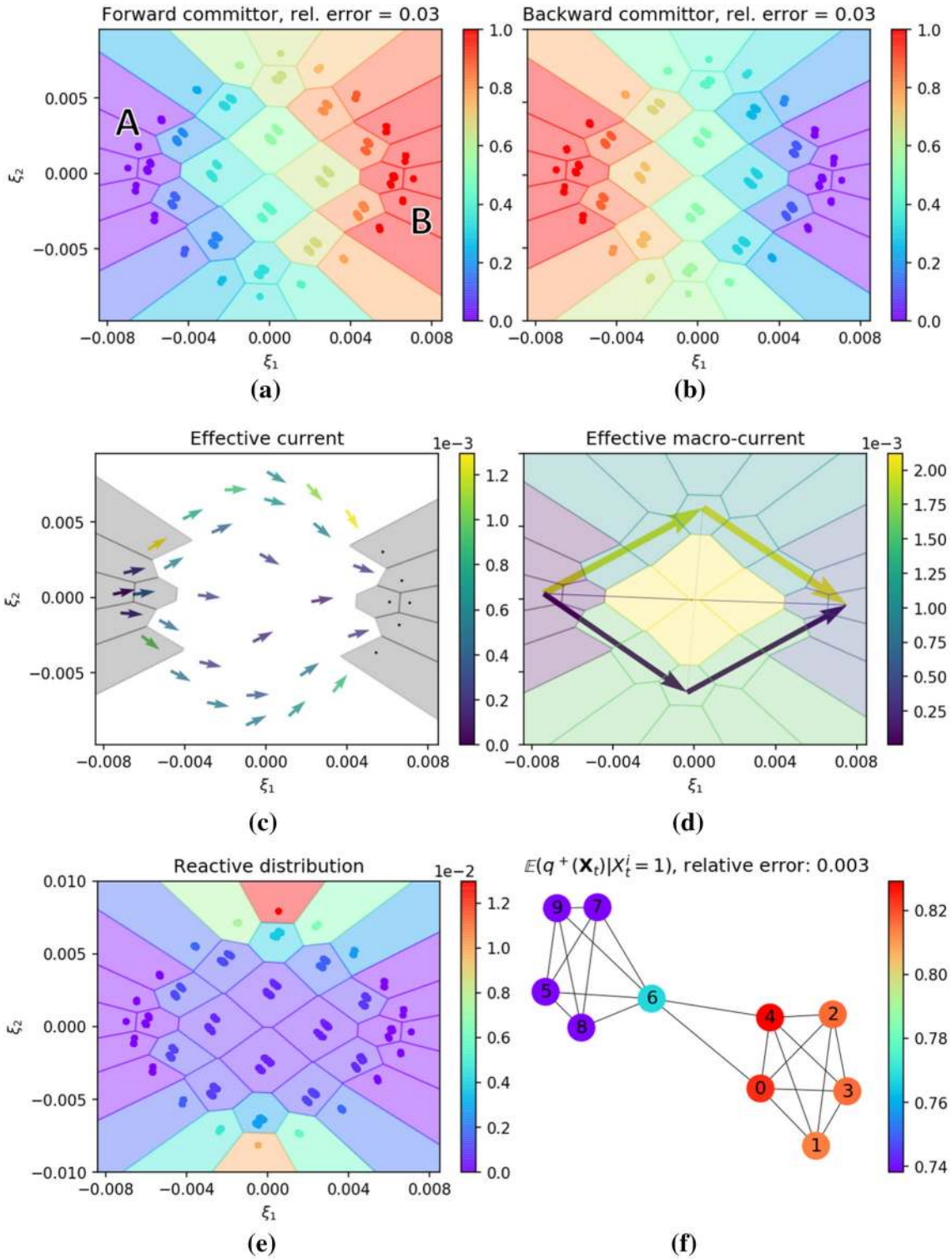

Fig. 8 Tipping analysis for *Example 1:* **a**, **b** Estimated committors on the discretized space. **c** Effective current, A and B are indicated by the two shaded areas. **d** Effective macro-current through the three channels indicated in shaded blue, yellow and green. **e** Reactive distribution. **f** Agents as indicators of the overall tipping

matrix on this discrete space by using short trajectory snippets of total length $T = 100,000$.

The results of the tipping analysis between $A = \{\leq 2$ agents are active$\}$ and $B = \{\geq 8$ agents are active$\}$ are presented in Fig. 8. In the panels **a** and **b** we show the committors on the reduced state space.

The original state space is small enough to be able to solve the system of linear equations (3) and (4) for the exact forward and backward committors, denoted below by $q^{\pm}_{\text{exact}}$. Thus we can study the relative error in the π-weighted l_2-norm between the estimated committors and the exact committors:

$$\frac{\sqrt{\sum_{\mathbf{x}\in\mathbb{X}} \left(q^{\pm}_{\text{exact}}(\mathbf{x}) - q^{\pm}(\mathbf{x})\right)^2 \pi(\mathbf{x})}}{\sqrt{\sum_{\mathbf{x}\in\mathbb{X}} \left(q^{\pm}_{\text{exact}}(\mathbf{x})\right)^2 \pi(\mathbf{x})}}.$$

The weighting with the stationary distribution π is the natural weighting for the dynamics. The relative error of both committors is 0.03 and confirms that the collective variables allow a good approximation of the tipping dynamics.

The forward committor is not perfectly symmetric with respect to the two blocks: when block 1 has completely tipped but block 2 has not (these are the states around $\xi_1 = 0$ and $\xi_2 > 0.005$, compare with Fig. 4c), the forward committor is much higher than in the opposite scenario, when block 2 has tipped but block 1 not (the states around $\xi_1 = 0$ and $\xi_2 < -0.005$). Also the reactive distribution is higher when block 1 has tipped and 2 has not. The transition rate amounts to $k^{AB} = 0.0039$ meaning that in a stationary trajectory, a transition from A to B of duration $t^{AB} = 18.85$ is started on average every $1/k^{AB} \approx 256$th time step. From the effective current[7] in Fig. 8e we can see that most of the transition flux from A to B goes along two pathways:

(I) $A \to$ agents in block 1 get active $\to$ agents in block 2 get active $\to B$
(II) $A \to$ agents in block 2 get active $\to$ agents in block 1 get active $\to B$.

To better compare the likelihood of both transition channels, we group the states of each channel together by hand, compare the coloring in Fig. 8f, and compute the reactive macro-currents F^{AB} and the effective macro-currents F^{+} through these channels. Transitions along channel (I) contribute 53% to k^{AB}, while channel (II) only contributes 41% to the rate. We can now confirm that there is more effective current going through channel (I). The reason should lie in the asymmetry of the network between block 1 and 2: Agents 0 and 4 of block 1 are both connected to agent 6, see the network in Fig. 4b. And from the ABM interaction rules, we can deduce that the likelihood that agent 0 and 4 become active when agent 6 is active is smaller than the likelihood that agent 6 becomes active after agents 0 and 4. These results also fit with the asymmetry in the committor: as soon as block 1 has become active, it is very likely that block 2 also becomes active.

To further study the role of each individual agent with respect to the overall tipping between A and B, we consider the expected forward committor conditioned on agent i being active. When the forward committor is conditioned on agent i being active, the agents with the largest

$$\mathbb{E}(q^{+}(X_t) \mid X_t^i = 1) =: I_i^{AB}$$

are the best (individual-agent) *indicators* that the overall tipping of the population will soon happen. When these agents are active, the system is the most likely to tip to B, thus one should especially consider these agents to access the tipping likelihood.

We can estimate I_i^{AB} by a Monte-Carlo approximation with a sufficiently long stationary ABM trajectory $(\mathbf{x}_t)_{t=1,\ldots,T}$:

$$\begin{aligned} I_i^{AB} &= \frac{\mathbb{E}\left(q^{+}(X_t)\,\mathbb{1}_{\{X_t^i=1\}}\right)}{\mathbb{P}\left(X_t^i = 1\right)} \\ &= \frac{\sum_{\mathbf{x}\in\mathbb{X}} q^{+}(\mathbf{x})\,\mathbb{1}_{\{x^i=1\}}(\mathbf{x})\,\pi(\mathbf{x})}{\sum_{\mathbf{x}\in\mathbb{X}} \mathbb{1}_{\{x^i=1\}}(\mathbf{x})\,\pi(\mathbf{x})} \\ &\approx \frac{\sum_{t=1}^{T} q^{+}(\tilde{\xi}(\mathbf{x}_t))\,\mathbb{1}_{\{x_t^i=1\}}(\mathbf{x}_t)}{\sum_{t=1}^{T} \mathbb{1}_{\{x_t^i=1\}}(\mathbf{x}_t)}, \end{aligned}$$

where we introduced the discrete collective variable $\tilde{\xi}(\mathbf{x}) = m$ whenever $\xi(\mathbf{x}) \in V_m$. From Fig. 8 we can see that the agents from block 1 are the better tipping indicators. Moreover, agents 0 and 4 are the best indicators of tipping, while agent 6 is the best indicator of block 2. This is probably due to them being connected to the other block, thus increasing the tipping likelihood when they are active. One has to be careful in the interpretation of I^{AB}, it only shows us the correlations of the state of agent i and the forward committor and not a causation, i.e., which agent has the largest individual impact on the overall tipping.

Example 2 continued: After discretizing the projected state space into 150 cells and estimating a transition matrix using trajectory snippets of total length $T = 40,000,000$, we show the tipping analysis for a population of four connected blocks in Fig. 9. As A we consider states where $\leq 25\%$ of agents are active, and as B the states where $\geq 75\%$ are active. The dominant Diffusion Maps coordinate encodes the number of agents that are active, and from Fig. 9a we can see that along this coordinate the forward committor increases in distinct steps from 0 to 1. Due to the faster decorrelation inside each metastable set, i.e., in the regions where agents in the same block are behaving conform, the committor is constant there. From the committors we computed

[7] The threshold model is only very slightly non-reversible, therefore we are not doing a flux-decomposition into cycles and productive parts and instead use the approximately cycle-free effective current.

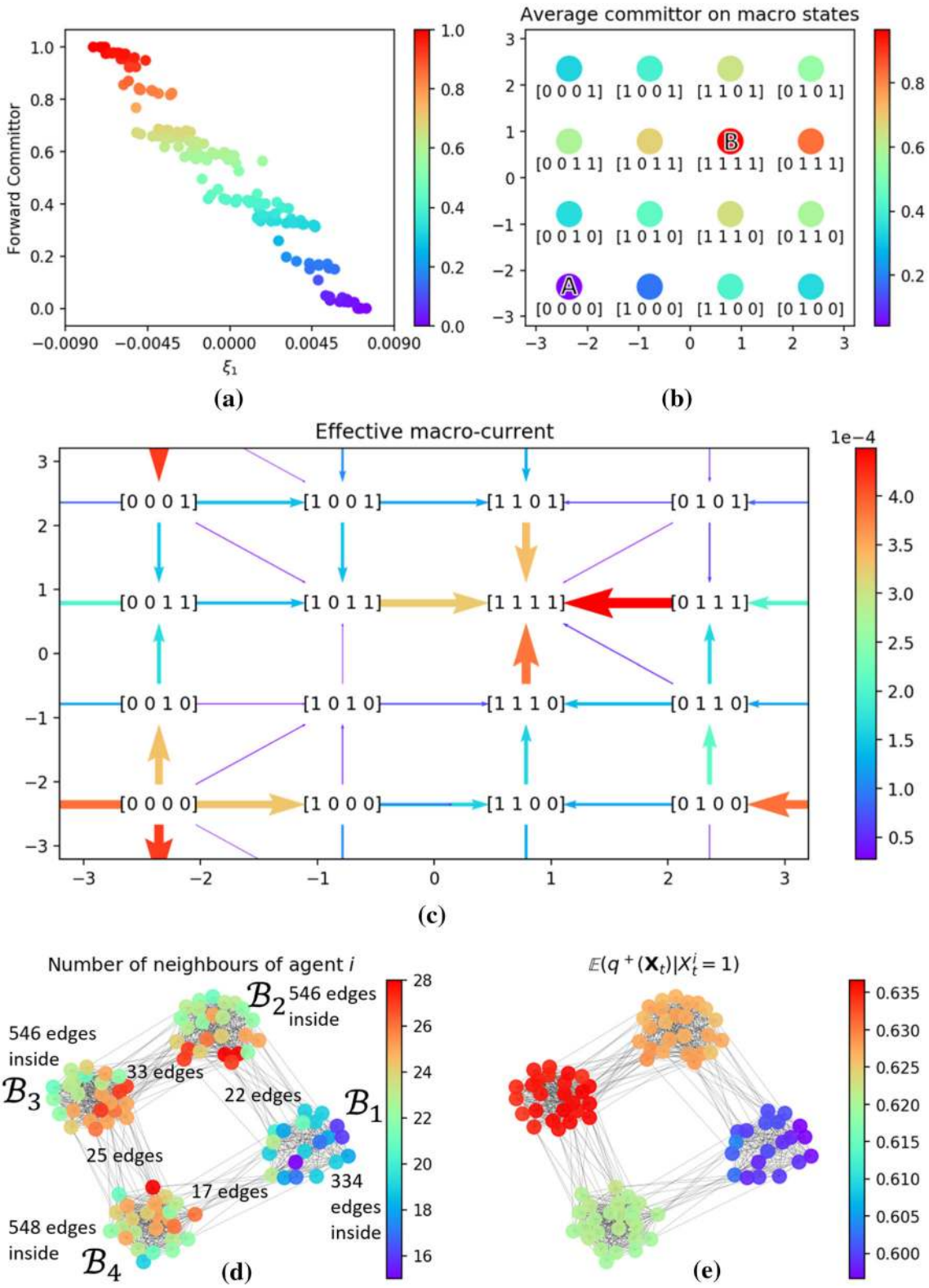

Fig. 9 Tipping analysis of *Example 2:* **a** Forward committor against the dominant Diffusion Maps coordinate. **b** Mean forward committor on the macrostates that are placed on a torus. We denoted macrostates as a 4-D vector of 0's and 1's decoding the majority activity in each of the four blocks, e.g., $[0, 0, 1, 0]$ reads as majority of agents in block 1,2 and 4 are inactive and majority in block 3 is active. **c** Effective macro-current, the color and width of the arrow indicates the magnitude of the current. **d** Number of neighbours of each agent as well as the total number of connections inside and between blocks. **e** Agents as indicators of overall tipping

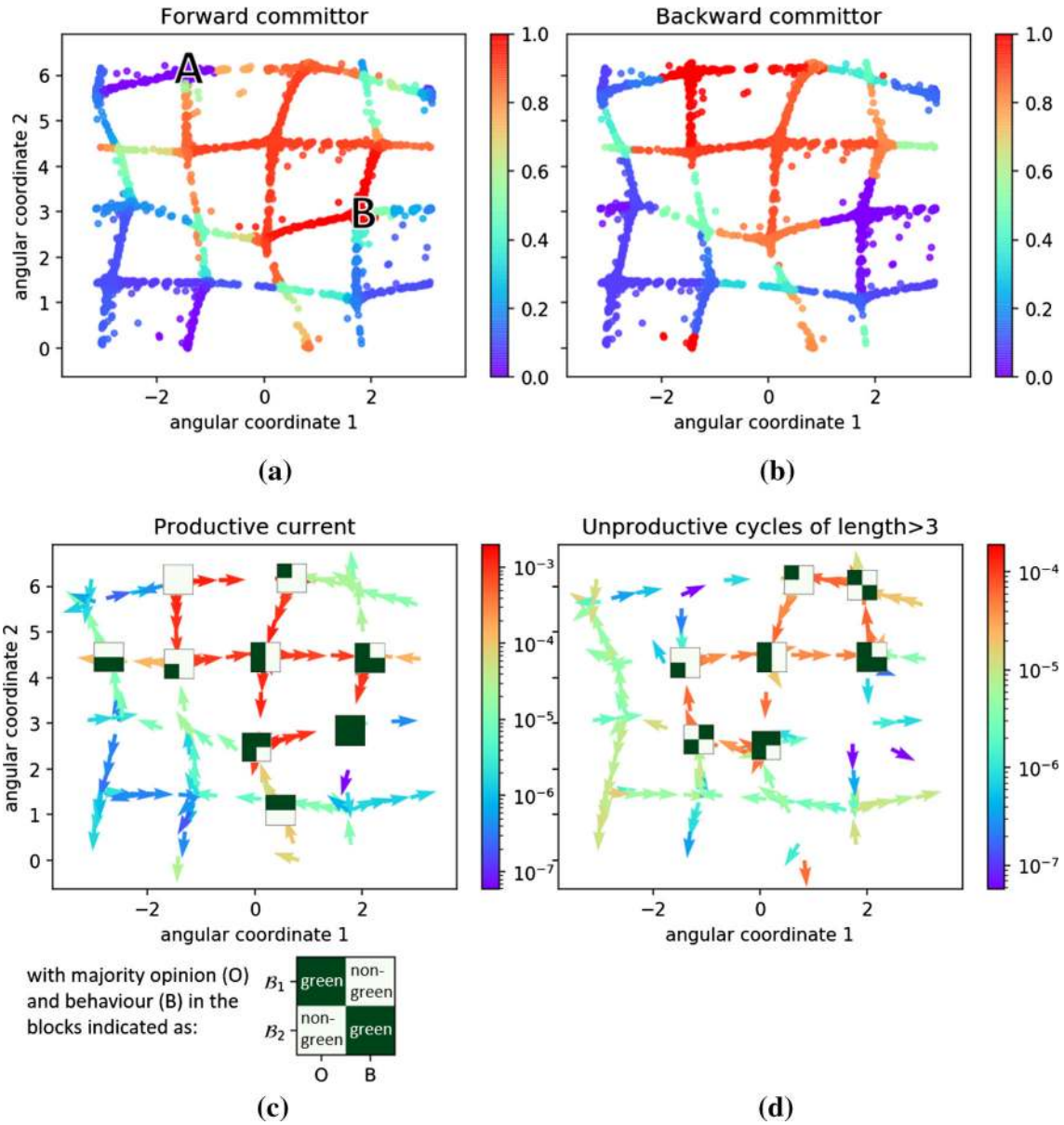

Fig. 10 Tipping analysis of *Example 3:* **a** Forward and **b** backward committor. **c** Productive, cycle-free current from A to B (note the logarithmic colour scale). **d** Unproductive current of cycles whose length is larger than 3. To get a clearer picture, we only plotted the flux produced by large cycles. For **c** and **d** we labeled the important macrostates by a table that indicates whether for that macrostate the majority in a block has a non-green or green opinion (O) or behaviour (B)

transition statistics, such as the average duration of reactive trajectories $t^{AB} = 79.3$ and their frequency: In a long stationary trajectory, a transition from A to B is completed on average every $1/k^{AB} \approx 627$th time step.

We again are interested in clustering states together to easier understand the transition dynamics and get a transition network. Since the system is much larger this time, we want to group cells together which are dynamically close by means of a clustering algorithm such as the well-known fuzzy clustering method PCCA+ [52] which stands for *Robust Perron Cluster Cluster Analysis*. We will use PCCA+ for non-reversible processes [14,16,51], implemented in [61], which takes the dominant real Schur vectors of the transition matrix and by a linear transformation maps them into a set of non-negative membership vectors that form a partition of unity and are as crisp as possible. Other optimization criteria also exist [52]. By assigning each point to the block with the highest membership value, we can make the clustering crisp. The advantage of PCCA+ compared to other fuzzy clustering algorithms is that it takes the dynamical information into account and results in a clustering that tries to preserve the slow

time scales of the dynamical process. Since we expect the macrostates to be of the form where the majority of agents in each block is either inactive or active, we cluster the states into 2^4 macrostates. These macrostates also correspond to the corners of the tesseract. The macrostates were then placed on a 2-D torus such that the transition network can easily be visualized.

In Fig. 9b, c we show the mean forward committor on the macrostates as well as the resulting transition network given by the effective macro-current. The macro-current is larger between macrostates where a neighbouring block tips than for a non-neighbouring block. Thus the dominant pathways from A to B are of the form of a tipping cascade from one block to its neighbours and then to their neighbours etc. The macro-current also indicates that it is most likely for block 4 to tip first and for block 1 to tip last. This can be explained as follows: Every agent in block 4 has on average 21.92 neighbours from the same block and 1.68 neighbours from the other blocks. Compared to the other blocks, agents from block 4 have the highest proportion of neighbours from the same block. Thus block 4 is the most independent block and therefore can change its activity most freely. The role of block 1 is also special. It is the smallest block with only 20 agents and also the block where each agent has the largest proportion of extraneous neighbours. The role of block 1 is also reflected in the mean forward committor values: Out of all the macrostates, where only one block has tipped, the committor is the smallest when only block 1 has tipped. This indicates that when block 1 has tipped, it easily tips back due to the strong influence from its neighbouring blocks. Moreover, out of all the macrostates where three blocks have tipped, the forward committor is the highest when block 1 is the still inactive block.

For the network of four blocks we can study which agents are the best indicators of the overall tipping, see Fig. 9e. We can immediately see that the values of I_i^{AB} do not differ that much for the different agents, possibly due to the four blocks being of a rather similar size and similarly connected. Still, block 3 seems to result in the highest expected forward committor when an agent of that block is active. Block 3 has the most connections to other blocks, and can therefore possibly exert the most influence on neighbouring blocks. This might explain why the expected forward committor is the largest when an agent from block 3 is active.

Complex contagion model In this model we are interested in analysing the tipping pathways between states where the majority has a non-green opinion and behaviour to states, where the majority has a green opinion and behaviour.

Example 3 continued: We discretized the projected space into 150 Voronoi cells and estimated the transition matrix on this space using 100,000 short trajectory snippets. The tipping analysis between the regions

$A =$
$\{\leq 20\%$ of the population have a green opinion and behaviour$\}$,

$B =$
$\{\geq 80\%$ of the population have a green opinion and behaviour$\}$

is shown in Fig. 10. To better understand the projected states, one can compare with the coloring in Fig. 6d or the indicated macro states in Fig. 10c, d. The two blocks behave as coupled oscillators that are mostly synchronized in a stationary regime. When the majority of agents in one block changes their behaviours or opinions in the cyclic fashion, the other block will likely follow. There is a strong direction in the dynamics, i.e. the dynamics most of the time follow the same path, thus the forward committor is close to deterministic for many states, i.e., takes values close to 0 and 1, see Fig. 10a. When the two blocks first change their opinions and then their behaviours from *non-green* to *green*, the forward committor is close to 1 and when they change their opinions and behaviours back to *non-green*, the forward committor is close to 0. But there are also some states with a committor around $\frac{1}{2}$ and thus the future states thereafter are less predictable. The backward committor is similarly very deterministic for a large part of the statespace. We will next look at the current to understand the possible transition pathways from A to B and understand what happens when the committors are close to $\frac{1}{2}$.

Due to high non-reversibility, the effective flux is no longer cycle-free. Instead we can decompose the reactive flux into a productive, cycle-free and an unproductive cyclic flux, see Fig. 10c, d. From the decomposition we see that the dominant productive pathways are of the form:

(I) $A \rightarrow$ agents in one of the blocks change their opinion to *green* $\rightarrow$ agents in the other block change their opinion to *green* $\rightarrow$ agents in one of the blocks change their behaviour to *green* $\rightarrow$ agents in the other block change their behaviour to *green* $\rightarrow B$,

while there are also some less likely productive paths:

(II) $A \rightarrow$ agents in one of the blocks change their opinion to *green* $\rightarrow$ agents in the same block change their behaviour to *green* $\rightarrow$ agents in the other block change their opinion to *green* $\rightarrow$ agents in the other block change their behaviour to *green* $\rightarrow B$.

The dominant unproductive cycles are of the general form:

(III) Both blocks have a *non-green* behaviour, majority of agents in block 1 (resp. 2) have a *green* opinion $\rightarrow$ agents in block 2 (resp. 1) change their opinion to *green* $\rightarrow$ agents in block 2 (resp. 1) change their behaviour to *green* $\rightarrow$ agents in block 2 (resp. 1) change their opinion back to *non-green* $\rightarrow$ then agents in block 2 (resp. 1) change their behaviour back to *non-green*.

In these unproductive cycles, one block does a solo-cycle through the behaviour and opinion space. These are also common in coupled oscillators and called "2π phase jumps" [1,49].

By comparing the strength of the flux along the dominant productive paths (I) (around $9\times10^{-4}-10^{-3}$) with the values of the current along the dominant unproductive cycles (III) (around $4\times10^{-5}-6\times10^{-5}$) in Fig. 10c, d, we can deduce that the pathways (I) are visited 15−25 times as much as the dominant cyclic structure (III).

Beyond the dominant pathways, we can give some general quantitative statements: conditioned on being on a reactive trajectory, the probability to be on a productive path is $(H^{AB})^{-1}\sum_{x,y} f^P(x,y) = \frac{0.05}{0.285} = 0.175$, while the probability to be on a cycle of length > 3 is $(H^{AB})^{-1}\sum_{x,y} f^{U,>3}(x,y) = \frac{0.004}{0.285} = 0.014$. The remaining conditional probability is attributed to cycles of length ≤ 3.

5 Conclusion

In this paper we showed how to quantitatively study noise-induced tipping pathways in high-dimensional, stationary models of heterogeneous agents. For complicated agent-based models, analytically deriving reduced equations, e.g., ODEs or SDEs, is no longer possible or one has to accept large approximation errors. Here we instead relied on simulations of the model to estimate a low-dimensional representation of the population states in terms of collective variables. In our two guiding models, agents are strongly affiliated with a subpopulation. Due to the local interaction rules, those population states, where the individual agents in the same subpopulation agree on their actions or attitudes, are metastable. The population states approximately lie on the skeleton of a hypercube that can be parametrized with just a few coordinates. The corners of the hypercube represent the metastable states while the edges make up the transition paths between metastable states. Thus the estimated reduced states can describe all the macro-scale patterns and large shifts and changes in the population of agents. In the two considered ABMs, tipping between the two extreme metastable regimes in the system happened as a tipping cascade among connected subpopulations. We applied Transition Path Theory to quantify the tipping dynamics and could for instance uncover the dominant cascading pathways as well as possible loop-dynamics on the way from A to B.

It is noteworthy to mention that TPT can quantify the tipping paths without relying on actual sampled transition paths. By estimating the transition matrix from short samples, the local information in the short samples can be combined to solve for the global committor functions. Still the simulation data needs to cover all the important parts of the model state space, i.e., the sets A and B as well as the visited states during transitions.

By studying which agent i results in the highest expected forward committor conditioned on them being in a certain state, I_i^{AB}, we could assess which agents in the network are the best indicators for tipping towards B.

Several open aspects and questions remain:

1. In order to better understand the quantity I_i^{AB}, one could systematically study I_i^{AB} for different small networks, similar as in [25,26], as well as compare it to different centrality measures for network nodes.
2. By studying other general types of ABM dynamics or interactions on non-modular networks, could one find other generic forms of the low-dimensional manifold on which the population states concentrate?
3. Another prospect would be the study of more realistic ABMs or dynamics on real-world networks.
4. In this paper we studied tipping in stationary ABMs where the tipping is only due to the noise facilitating rare transitions. But as mentioned in the introduction, agent-based models are often not stationary. Therefore at next it would be important to also consider tipping in non-stationary ABMs, e.g., ABMs influenced by some external parameter variations [22].

Acknowledgements We would like to thank Marc Wiedermann for discussions about Granovetter's threshold model and Alexander Sikorski for helping with speeding-up the ABM simulation code and many discussions. We are grateful to Marvin Lücke for carefully reading the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. Luzie Helfmann acknowledges support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—The BerlinMathematics Research CenterMATH+ (EXC-2046/1, project ID: 390685689).

Author contribution statement

JH, PK, JK and CS supervised the project, LH performed numerical computations and analyzed the results. All authors discussed the results and contributed to the final manuscript.

Data Availability Statement This manuscript has no associated data or the data will not be deposited. [Authors' comment: The datasets analysed in this paper can be generated from the agent-based models in Sect. 2.]

References

1. A. Arenas, A. Díaz-Guilera, J. Kurths, Y. Moreno, C. Zhou, Synchronization in complex networks. Phys. Rep. **469**(3), 93–153 (2008)
2. P. Ashwin, S. Wieczorek, R. Vitolo, P. Cox, Tipping points in open systems: bifurcation, noise-induced and rate-dependent examples in the climate system. Philos. Trans. R Soc. A Math. Phys. Eng. Sci. **370**(1962), 1166–1184 (2012)
3. R. Banisch, N.D. Conrad, Cycle-flow-based module detection in directed recurrence networks. Europhys. Lett. **108**(6), 68008 (2015)
4. R. Banisch, N. Djurdjevac, C. Schütte, Reactive flows and unproductive cycles in irreversible Markov chains. Eur. Phys. J. Spec. Top. **224**(12), 2369–2387 (2015)
5. F. Battiston, G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J.-G. Young, G. Petri, Networks beyond pairwise interactions: structure and dynamics. Phys. Rep. **874**, 1–92 (2020)
6. T. Berry, J. Harlim, Variable bandwidth diffusion kernels. Appl. Comput. Harmon. Anal. **40**(1), 68–96 (2016)
7. Y. Cai, T.M. Lenton, T.S. Lontzek, Risk of multiple interacting tipping points should encourage rapid CO2 emission reduction. Nat. Clim. Change **6**(5), 520–525 (2016)
8. M. Cameron, E. Vanden-Eijnden, Flows in complex networks: theory, algorithms, and application to Lennard–Jones cluster rearrangement. J. Stat. Phys. **156**(3), 427–454 (2014)
9. D. Centola, M. Macy, Complex contagions and the weakness of long ties. Am. J. Sociol. **113**(3), 702–734 (2007)
10. P. Cilliers, Boundaries, hierarchies and networks in complex systems. Int. J. Innov. Manag. **5**(02), 135–147 (2001)
11. P. Cilliers, D. Spurrett, Complexity and postmodernism: understanding complex systems. South Afr. J. Philos. **18**(2), 258–274 (1999)
12. R.R. Coifman, S. Lafon, Diffusion Maps. Appl. Comput. Harmon. Anal. **21**(1), 5–30 (2006)
13. R.R. Coifman, S. Lafon, Geometric harmonics: a novel tool for multiscale out-of-sample extension of empirical functions. Appl. Comput. Harmon. Anal. **21**(1), 31–52 (2006)
14. N.D. Conrad, M. Weber, C. Schütte, Finding dominant structures of nonreversible Markov processes. Multiscale Model. Simul. **14**(4), 1319–1340 (2016)
15. G. Deffuant, D. Neau, F. Amblard, G. Weisbuch, Mixing beliefs among interacting agents. Adv. Complex Syst. **3**(01n04), 87–98 (2000)
16. K. Fackeldey, A. Sikorski, M. Weber, Spectral clustering for non-reversible Markov chains. Comput. Appl. Math. **37**(5), 6376–6391 (2018)
17. J. Finkel, D.S. Abbot, J. Weare, Path properties of atmospheric transitions: illustration with a low-order sudden stratospheric warming model. J. Atmos. Sci. **77**(7), 2327–2347 (2020)
18. J. Finkel, R. J. Webber, D. S. Abbot, E. P. Gerber, J. Weare. Learning forecasts of rare stratospheric transitions from short simulations. arXiv:2102.07760 (2021)
19. M. Girvan, M.E.J. Newman, Community structure in social and biological networks. Proc. Natl. Acad. Sci. **99**(12), 7821–7826 (2002)
20. M. Gladwell, *The tipping point: how little things can make a big difference* (Little, Brown, 2006)
21. M. Granovetter, Threshold models of collective behavior. Am. J. Sociol. **83**(6), 1420–1443 (1978)
22. L. Helfmann, E.R. Borrell, C. Schütte, P. Koltai, Extending transition path theory: periodically driven and finite-time dynamics. J. Nonlinear Sci. **30**(6), 3321–3366 (2020)
23. L. Helfmann, E. R. Borrell. Pytpt. https://github.com/LuzieH/pytpt. Accessed 13 May 2021
24. R.A. Holley, T.M Liggett. Ergodic theorems for weakly interacting infinite systems and the voter model. Ann. Probab. **3**, 643–663 (1975)
25. P. Holme, Three faces of node importance in network epidemiology: exact results for small graphs. Phys. Rev. E **96**(6), 062305 (2017)
26. P. Holme, L. Tupikina, Epidemic extinction in networks: insights from the 12 110 smallest graphs. New J. Phys. **20**(11), 113042 (2018)
27. L.R. Izquierdo, S.S. Izquierdo, J.M. Galan, J.I. Santos, Techniques to understand computer simulations: Markov chain analysis. J. Artif. Soc. Soc. Simul. **12**(1), 6 (2009)
28. D.-Q. Jiang, D. Jiang, M. Qian. *Mathematical Theory of Nonequilibrium Steady States: On the Frontier of Probability and Dynamical Systems*, vol. 1833. (Springer, 2004)
29. S.L. Kalpazidou. *Cycle Representations of Markov Processes*, vol. 28. (Springer, 2007)
30. F.P. Kemeth, T. Bertalan, T. Thiem, F. Dietrich, S.J. Moon, C.R. Laing, I.G. Kevrekidis. Learning emergent PDEs in a learned emergent space. arXiv:2012.12738 (2020)
31. Y. Khoo, L. Jianfeng, L. Ying, Solving for high-dimensional committor functions using artificial neural networks. Res. Math. Sci. **6**(1), 1–13 (2019)
32. P. Koltai, S. Weiss, Diffusion Maps embedding and transition matrix analysis of the large-scale flow structure in turbulent Rayleigh-Bénard convection. Nonlinearity **33**(4), 1723 (2020)
33. E. Kriegler, J.W. Hall, H. Held, R. Dawson, H.J. Schellnhuber, Imprecise probability assessment of tipping points in the climate system. Proc. Natl. Acad. Sci. **106**(13), 5041–5046 (2009)
34. H. Li, Y. Khoo, Y. Ren, L. Ying. Solving for high dimensional committor functions using neural network with online approximation to derivatives. arXiv:2012.06727 (2020)

35. Q. Li, B. Lin, W. Ren, Computing committor functions for the study of rare events using deep learning. J. Chem. Phys. **151**(5), 054112 (2019)
36. M. Lindner, F. Hellmann, Stochastic basins of attraction and generalized committor functions. Phys. Rev. E **100**(2), 022124 (2019)
37. P. Liu, H.R. Safford, I.D. Couzin, I.G. Kevrekidis, Coarse-grained variables for particle-based models: Diffusion Maps and animal swarming simulations. Comput. Part. Mech. **1**(4), 425–440 (2014)
38. D. Lucente, S. Duffner, C. Herbert, J. Rolland, F. Bouchet. Machine learning of committor functions for predicting high impact climate events. arXiv:1910.11736 (2019)
39. M.W. Macy, R. Willer, From factors to actors: computational sociology and agent-based modeling. Annu. Rev. Sociol. **28**(1), 143–166 (2002)
40. C. Marschler, J. Starke, P. Liu, I.G. Kevrekidis, Coarse-grained particle model for pedestrian flow using Diffusion Maps. Phys. Rev. E **89**(1), 013304 (2014)
41. J.-D. Mathias, J.M. Anderies, J. Baggio, J. Hodbod, S. Huet, M.A. Janssen, M. Milkoreit, M. Schoon, Exploring non-linear transition pathways in social-ecological systems. Sci. Rep. **10**(1), 1–12 (2020)
42. P. Metzner, F. Noé, C. Schütte, Estimating the sampling error: distribution of transition matrices and functions of transition matrices for given trajectory data. Phys. Rev. E **80**(2), 021106 (2009)
43. P. Metzner, C. Schütte, E. Vanden-Eijnden, Transition path theory for Markov jump processes. Multiscale Model. Simul. **7**(3), 1192–1219 (2009)
44. P. Miron, F. J. Beron-Vera, L. Helfmann, P. Koltai. Transition paths of marine debris and the stability of the garbage patches. Chaos **31**, 033101 (2021)
45. J.-H. Niemann, S. Winkelmann, S. Wolf, C. Schütte, Population limits and large timescales. Agent-Based Model. **31**, 033140 (2020)
46. F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, T.R. Weikl, Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. Proc. Natl. Acad. Sci. **106**(45), 19011–19016 (2009)
47. J.R. Norris. *Markov Chains. Cambridge Series in Statistical and Probabilistic Mathematics* (Cambridge University Press, 1997)
48. K. Nyborg, J. M. Anderies, A. Dannenberg, T. Lindahl, C. Schill, M. Schlüter, W. N. Adger, K. J. Arrow, S. Barrett, S. Carpenter, et al., Social norms as solutions. Science **354**(6308), 42–43 (2016)
49. A. Pikovsky, J. Kurths, M. Rosenblum, J. Kurths. *Synchronization: a universal concept in nonlinear sciences. Cambridge Nonlinear Science Series*, No. 12 (Cambridge University Press, 2003)
50. M.A. Porter, J.P. Gleeson, *Dynamical systems on networks*. Frontiers in Applied Dynamical Systems: Reviews and Tutorials, vol. 4 (2016)
51. B. Reuter, K. Fackeldey, M. Weber, Generalized Markov modeling of nonreversible molecular kinetics. J. Chem. Phys. **150**(17), 174103 (2019)
52. S. Röblitz, M. Weber, Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification. Adv. Data Anal. Classif. **7**(2), 147–179 (2013)
53. M.A. Rohrdanz, W. Zheng, C. Clementi, Discovering mountain passes via torchlight: methods for the definition of reaction coordinates and pathways in complex macromolecular reactions. Annu. Rev. Phys. Chem. **64**, 295–316 (2013)
54. S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
55. S. Sahasrabudhe, A.E. Motter, Rescuing ecosystems from extinction cascades through compensatory perturbations. Nat. Commun. **2**(1), 1–8 (2011)
56. C. Schütte, M. Sarich. *Metastability and Markov State Models in Molecular Dynamics*, vol. 24. American Mathematical Society (2013)
57. L. Serdukova, Y. Zheng, J. Duan, J. Kurths, Stochastic basins of attraction for metastable states. Chaos **26**(7), 073117 (2016)
58. L. Serdukova, Y. Zheng, J. Duan, J. Kurths, Metastability for discontinuous dynamical systems under Lévy noise: case study on Amazonian vegetation. Sci. Rep. **7**(1), 1–13 (2017)
59. S. Sharpe, T.M. Lenton. Upward-scaling tipping cascades to meet climate goals–plausible grounds for hope. UCL Institute for Innovation and Public Purpose, Working Paper Series (2020)
60. C.I. Siettos, C.W. Gear, I.G. Kevrekidis, An equation-free approach to agent-based computation: bifurcation analysis and control of stationary states. Europhys. Lett. **99**(4), 48007 (2012)
61. A. Sikorski. cmdtools. https://github.com/zib-cmd/cmdtools. Accessed 13 May 2021
62. H.A. Simon, The architecture of complexity. In *Facets of Systems Science*, pp. 457–476 (Springer, 1991)
63. A. Sîrbu, V. Loreto, V.D.P. Servedio, F. Tria. Opinion dynamics: models, extensions and external effects. In *Participatory Sensing, Opinions and Collective Qwareness*, pp. 363–401 (Springer, 2017)
64. J.B. Tenenbaum, V. De Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction. Science **290**(5500), 2319–2323 (2000)
65. E.H. Thiede, D. Giannakis, A.R. Dinner, J. Weare, Galerkin approximation of dynamical quantities using trajectory data. J. Chem. Phys. **150**(24), 244111 (2019)
66. A.C. Tsoumanis, C.I. Siettos, G.V. Bafas, I.G. Kevrekidis, Equation-free multiscale computations in social networks: from agent-based modeling to coarse-grained stability and bifurcation analysis. Int J Bifurc Chaos **20**(11), 3673–3688 (2010)
67. J. Ugander, B. Karrer, L. Backstrom, C. Marlow, The anatomy of the facebook social graph. arXiv:1111.4503 (2011)
68. S.M. Ulam, *A Collection of Mathematical Problems* (Interscience Publisher, 1960)
69. L. Van der Maaten, G. Hinton, Visualizing data using t-sne. J. Mach. Learn. Res. **9**(11) (2008)
70. E. Vanden-Eijnden. Transition path theory. In: *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology* ed. by M. Ferrario, G. Ciccotti, K. Binder, vol. 1, pp. 453–493 (Springer, Berlin, 2006)
71. E. Vanden-Eijnden. Transition path theory. In *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation* ed. by G. R.

Bowman, V. S. Pande, F. Noé, chapter 7, pp. 91–100 (Springer, 2013)
72. D.J. Watts, A simple model of global cascades on random networks. Proc. Natl. Acad. Sci. **99**(9), 5766–5771 (2002)
73. E. Weinan, E. Vanden-Eijnden. Towards a theory of transition paths. J. Stat. Phys. **123**(3), 503–523 (2006)
74. M. Wiedermann, E. K. Smith, J. Heitzig, J. F. Donges. A network-based microfoundation of Granovetter's threshold model for social tipping. Sci. Rep. **10**(1), 1–10 (2020)
75. R. Winkelmann, J. F. Donges, E. K. Smith, M. Milkoreit, C. Eder, J. Heitzig, A. Katsanidou, M. Wiedermann, N. Wunderling, T. M. Lenton. Social tipping processes for sustainability: an analytical framework. arXiv:2010.04488 (2020)