

State Aggregations in Markov Chains and Block Models of Networks

Mauro Faccin¹, Michael T. Schaub^{2,3}, and Jean-Charles Delvenne^{1,4}

¹ICTEAM, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium

²Department of Engineering Science, University of Oxford, Oxford OX1 2JD, United Kingdom

³Department of Computer Science, RWTH Aachen University, 52074 Aachen, Germany

⁴CORE, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium

 (Received 4 May 2020; revised 17 June 2021; accepted 15 July 2021; published 12 August 2021)

We consider state-aggregation schemes for Markov chains from an information-theoretic perspective. Specifically, we consider aggregating the states of a Markov chain such that the mutual information of the aggregated states separated by T time steps is maximized. We show that for $T = 1$ this recovers the maximum-likelihood estimator of the degree-corrected stochastic block model as a particular case, which enables us to explain certain features of the likelihood landscape of this generative network model from a dynamical lens. We further highlight how we can uncover coherent, long-range dynamical modules for which considering a timescale $T \gg 1$ is essential. We demonstrate our results using synthetic flows and real-world ocean currents, where we are able to recover the fundamental features of the surface currents of the oceans.

DOI: [10.1103/PhysRevLett.127.078301](https://doi.org/10.1103/PhysRevLett.127.078301)

Systems comprising the interactions of many entities often exhibit complex dynamics that unfold within a large state space. A powerful idea to tame this complexity is to project the system state x_t at each time t onto a significantly smaller space and replace the original dynamics, say, of the form $x_{t+1} = f(x_t, x_{t-1}, \dots)$, with the simpler dynamics $y_{t+1} = g(y_t, y_{t-1}, \dots)$ of the projected state y_t . Such techniques abound in physics and other fields under headings such as model order reduction, coarse graining, variable or state aggregation, mode decomposition, or dimensionality reduction [1–11].

The success of these methods hinges on the choice of a projection $y_t = h(x_t)$ that retains the salient features of the original dynamics. For example, for a linear dynamics, a small subspace spanned by its dominant, low-frequency eigenmodes governs the long-term behavior. The neglected eigenmodes correspond to high-frequency modes describing short-lived transients. Projecting x_t onto the slow eigenmodes yields a system description y_t with theoretical guarantees on the reconstruction error of the original dynamics [8,11,12]. Accordingly, spectral techniques such as generalized Perron cluster analysis (GenPCCA) [13], which extract the dominant subspaces of a dynamics, have been proposed to address the problem of state aggregation. In other situations, we may also prefer to extract nondominant eigenvectors corresponding to medium or fast timescales [14–16].

Here, we consider a stationary Markov process on a discrete state space \mathcal{X} and explore information-theoretic strategies to find state aggregations that are akin to a nonlinear version of choosing between the slow and fast frequency modes. Given a state aggregation $y_t = h(x_t)$, we

study the time-lagged mutual information \mathcal{I}_T between the new state variables y_t and y_{t+T} for any timescale T . We call \mathcal{I}_T the autoinformation of the state aggregation scheme. Related information-theoretic ideas include influential works such as the information bottleneck method [17], approaches from computational mechanics [18], or the map equation [9] (see the Supplemental Material (SM) [19] for a discussion of related methods).

We demonstrate that our approach offers a fresh perspective on the problem of state aggregation. Specifically, we show that maximizing the autoinformation for unit timescales ($T = 1$) is, under certain conditions, equivalent to maximizing the likelihood of a degree-corrected stochastic block model (DCSBM) [45,46], a popular technique to recover community structure in networks [47–49]. Leveraging our dynamical perspective, we can thus pinpoint problems inherent to assumptions underlying the DCSBM. We further show how the time parameter T of the autoinformation leads to a nonlinear transformation mitigating these problems. Our scheme is thus particularly relevant for the analysis of trajectory data with trends emerging over longer timescales, which we illustrate by analyzing an ocean drifter dataset where we can reveal dominant patterns such as ocean currents over long timescales.

Autoinformation between aggregated states.—Consider a state aggregation $y_t = h(x_t)$ that maps the discrete state $x_t \in \mathcal{X}$ from a space of cardinality $|\mathcal{X}| = N$ onto a new state $y_t \in \mathcal{Y}$ in a smaller space of size $|\mathcal{Y}| = K \leq N$. This induces a partition of \mathcal{X} into “aggregation classes”: sets of states in \mathcal{X} mapped to the same aggregated state in \mathcal{Y} . Applying the mapping h to each observed state x_t of the

original trajectory yields a new trajectory that can be described by a stochastic dynamical system $y_{t+1} = g(y_t, y_{[t-1:-\infty]})$. Here, the symbol $y_{[\tau_1, \tau_2]}$ denotes the sequence of states $y_{\tau_1}, \dots, y_{\tau_2}$ from τ_1 until τ_2 .

To find an aggregation $y_t = h(x_t)$ whose states are informative about the evolution of the dynamics at the next time step, we seek a mapping h for which the mutual information $I(y_{t+1}, y_t)$ is as high as possible. It involves two terms of opposite signs:

$$I(y_{t+1}, y_t) = I(y_{t+1}; y_{[t, -\infty]}) - I(y_{t+1}; y_{[t-1, -\infty]} | y_t). \quad (1)$$

Maximizing $I(y_{t+1}; y_{[t, -\infty]})$ favors state aggregations that are as deterministic (or predictable) as possible. Minimizing $I(y_{t+1}; y_{[t-1, -\infty]} | y_t)$, however, leads to aggregations that are as Markovian as possible. Indeed, this term quantifies how much y_t deviates from a Markov process [50]: it is zero for a Markov process and positive otherwise. Note that even if x_t is a Markov process, the aggregated system $y_t = h(x_t)$ is not Markov in general; it is Markov if and only if the aggregation classes form a so-called lumpable partition of the transition matrix; see the SM [19].

We view Eq. (1) as a nonlinear counterpart to the unit time-lag linear autocorrelation of real-valued time series, which is pivotal for analyzing observables of linear dynamical systems, e.g., in signal processing or in the context of the fluctuation-dissipation theorem. Therefore, we call $I(y_{t+1}; y_t)$ the one-step autoinformation of the aggregated process. By the same rationale, we define the (T -step) autoinformation of the state aggregation h as

$$\mathcal{I}_T(h) := I(h(x_{t+T}); h(x_t)) = I(y_{t+T}; y_t) \quad (2a)$$

$$= H(y_t) - H(y_{t+T} | y_t), \quad (2b)$$

where $H(y_t) = H(y_{t+T})$ is the Shannon entropy of the aggregated state variables. Writing the autoinformation as the difference of conditional entropies highlights that it is maximized by an aggregated Markov chain with (i) a high number of approximately equiprobable states that maximize $H(y_t)$ and (ii) a low uncertainty $H(y_{t+T} | y_t)$ associated with the prediction of y_{t+T} based on state y_t .

Maximizing autoinformation as state-aggregation scheme.—The above discussion suggests maximizing the autoinformation $\mathcal{I}_T(h)$ over all possible state aggregations h as a possible scheme to obtain a reduced order description. Let us first explore the case in which we are given a desired cardinality K of the aggregated state space \mathcal{Y} , i.e., we look for a partition of \mathcal{X} into K aggregation classes. Denoting the space of all possible mappings to K states as \mathcal{H}_K , we arrive at the following optimization problem to obtain a state aggregation \hat{h}_T :

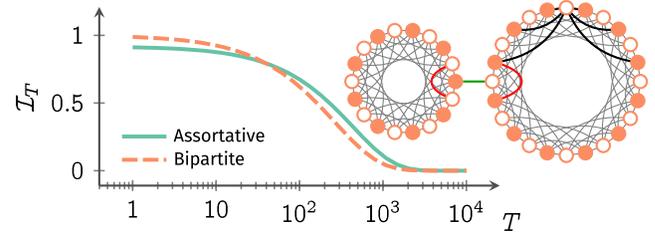


FIG. 1. Transition graph of a Markov chain with two alternative aggregations (inset schematic not of original size): an “assortative” split into two almost regular cyclic structures and a “disassortative,” almost bipartite split. The black edges exemplify the linking pattern. Two additional edges (red) break the bipartite symmetry and one joins the two cycles (green). The autoinformation results are shown for $N = 120 + 240$ nodes with average degree $\langle k \rangle = 10.02$. At short and long timescales, autoinformation is optimized in the bipartite or the assortative partition, respectively.

$$\hat{h}_T = \arg \max_{h \in \mathcal{H}_K} \mathcal{I}_T(h) = \arg \max_{h \in \mathcal{H}_K} I(y_{t+T}; y_t). \quad (3)$$

To gain intuition, consider Eq. (3) when x_t is a simple random walk process on an unweighted, undirected graph. Then, finding an optimal state aggregation is equivalent to finding an optimal partition of the nodes.

Figure 1 displays a simple state transition graph of a Markov chain with two cyclelike subparts connected by a single link. The cycles have even length and are constructed such that the graph is also almost bipartite. Let us now consider the problem of finding a state aggregation of this chain in $K = 2$ classes using autoinformation. The autoinformation associated with both aggregation classes is qualitatively similar: at each time step, the walker will likely both (i) change node type with respect to the (almost) bipartite structure and (ii) stay in the same cyclic structure. At short timescales, $H(y_{t+T} | y_t) \approx 0$ for both structures, and the $H(y_t)$ term in Eq. (2b) dominates. Accordingly, the bipartite partition, with slightly higher $H(y_t) \approx 1$, is preferred. For longer timescales, however, the second term of Eq. (2b) dominates, and the two-cycle partition is preferred: there is a smaller probability of leaving each cycle than of changing the bipartite aggregation class (see Fig. 1).

Relationships to the degree corrected stochastic block model.—A direct computation shows that optimizing Eq. (3) for $T = 1$ is (coincidentally) equivalent to solving a maximum-likelihood estimation problem for the DCSBM with K classes [45,46]. More precisely, $\hat{h}_{T=1} = \arg \max \ell_{\text{DCSBM}}(\mathbf{A})$, where $\mathbf{A} = [A_{ij}] \in \{0, 1\}^{N \times N}$ is the binary adjacency matrix of the graph and ℓ_{DCSBM} is the log-likelihood function of the DCSBM with model parameters given by their maximum-likelihood estimates (for a formal proof, see the SM).

The above result emphasizes that only paths of length 1 (edges) are essential to the likelihood function of the DCSBM, which derives from the mutual independence

of edges in a DCSBM. Interpreting the maximum-likelihood estimation for the DCSBM dynamically in terms of the autoinformation highlights this as a potential problem when fitting DCSBMs to graphs with long-range path structures. Indeed, since optimizing autoinformation for $T = 1$ for $K = 2$ amounts to fitting a two-group DCSBM, Fig. 1 shows that the two-cycle split would be missed when fitting such a graph via a DCSBM. Our dynamical standpoint sheds light on the underlying issue: when only considering trajectories of length 1, the description in terms of the bipartite structure will be preferred because it offers a more balanced partition of the states into two equiprobable classes. The specific path structure of this graph leads to a slow mixing of the chain within and between the two cycles, and the assortative split is thus not apparent at $T = 1$. Stated differently, at timescale $T = 1$, the bipartite switching is the dominant dynamical behavior of the Markov chain and fitting a DCSBM to the state-transition graph *correctly* captures this.

The importance of timescales.—Our approach also offers a way out of the above encountered dilemma: using $T \gg 1$ shifts the focus to the slow modes of the dynamics, for which the assortative split into the two cyclic structures becomes clear. The time parameter thus tailors the search to partitions that are dynamically relevant over longer timescales. Unlike with many community detection methods featuring a resolution parameter, the time parameter does not offset a null model linearly, but acts *nonlinearly* [see the SM, where we also prove the additional result that the optimal split into $K = 2$ equiprobable aggregation classes of *any* Markov chain tends to be either almost block diagonal (assortative) or almost bipartite (disassortative)].

How many aggregation classes?—In many scenarios, the number of aggregated states K can be gleaned from prior knowledge, and we thus have not discussed determining K . In a scenario where K is unknown, one would be tempted to optimize the autoinformation over all partitions without a constraint on K , but this would yield the trivial state aggregation $y_t = x_t$ (see the SM). This can be interpreted as data “overfitting”: without constraints on K , the best aggregation corresponds to the original model, which trivially captures all available information. To yield an aggregated description of size $K \leq N$ when maximizing the autoinformation, we have to impose additional constraints on the state-aggregation mapping h .

For a given quality criterion such as the autoinformation, two approaches are typically considered. One would be to find state-aggregation mappings via Eq. (3) for a varying number of states $K \in \{1, \dots, N\}$ and then select from among those solutions, e.g., using an elbow criterion (see the SM). Here, we follow another common approach by adding a complexity penalty to the objective function considered in Eq. (3). Optimizing the corresponding variational problem over all state-aggregation mappings leads to an aggregated system that maximizes

autoinformation while maintaining small complexity. This general approach can be interpreted in terms of Occam’s razor or a minimum description length (MDL) principle [51].

For simplicity, we choose the description length necessary to describe the aggregated states of the aggregated state space as a penalty term. Specifically, we consider the regularized autoinformation with an entropy penalty

$$\mathcal{I}_{\beta,T}(h) = \mathcal{I}_T(h) - \beta H(h(x_t)), \quad (4)$$

where β is a Lagrange multiplier for the regularization term (see the SM for a discussion of these parameters). However, our scheme is not bound to this specific complexity penalty, and other regularization schemes such as the Akaike information criterion [52] or ideas from Bayesian statistics and MDL-based modeling [51,53] may be considered. The specific choice of entropy for capturing the complexity of the partition can be seen as a smooth generalization of K , the number of classes, since K equally likely blocks translate into an entropy of $\log K$, while the entropy is also able to account for the size distribution of classes.

Like most combinatorial optimization problems, finding the aggregation that maximizes Eq. (4) is computationally difficult, and we thus have to resort to a heuristic optimization. Here, we use an ϵ -greedy optimization scheme akin to simulated annealing: starting from an initial partition, we stochastically loop over nodes and try to aggregate them with another class. If the regularized autoinformation improves, we aggregate the node with the new class with probability 1; otherwise, we aggregate with probability $e \propto e^{-\Delta \mathcal{I}_{\beta,T}/\tau}$, with τ a temperaturelike parameter that decreases along the maximization. A detailed discussion is given in the SM, and a reference implementation is publicly available [54].

Dynamical modules at short and long timescales.—The regularized autoinformation primarily provides a tool for state aggregations in Markov chains and dynamical data. However, due to its connection with maximum-likelihood estimation of a DCSBM (for $T = 1$), the regularized autoinformation also provides a dynamical view of certain model selection aspects under the DCSBM.

For concreteness, consider a random walk on a symmetric circular structure as the cycle of N nodes connected to the k nearest neighbors of Fig. 2. For short timescales, it is sensible for dynamical model reduction to aggregate small patches of the cycle that are unlikely to be left by the walker after T steps into aggregated states: the predictive power of such a fine-grained description outweighs the cost of the regularization term for most nonzero values of β . In particular, observe that maximizing Eq. (4) with $T = 1$ leads to a nontrivial number of aggregated states, as shown in Fig. 2. By symmetry arguments, which patches of the cycle we use as aggregated states is irrelevant, and there is a large number of equivalent optimal aggregated system

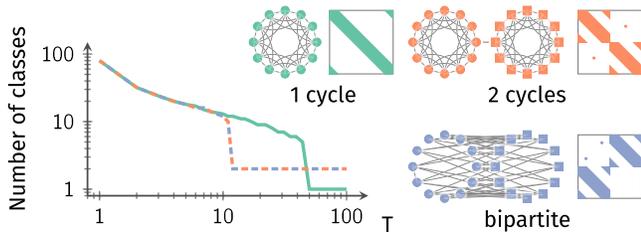


FIG. 2. Markov chains with natural timescales: a k -nearest-neighbor cycle with N nodes (green), two cycles of $N/2$ nodes connected by a single edge (orange), and a bipartite graph with a single link breaking the symmetry (blue); see insets for schematics. The plots correspond to graphs with 360 nodes with average degree $\langle k \rangle \approx 36$. The color of each line encodes the corresponding graph topology. At short timescales, the maximization of the regularized autoinformation ($\beta = 0.1$) tends to overfit the structure of these graphs with dense diagonal blocks (similar results hold for many community detection methods; see text). When increasing T , the algorithm finds the solution with the expected number of classes.

descriptions corresponding to different (symmetric, regular) partitions of the cycle.

Interestingly, qualitatively similar results hold *irrespective* of the regularization scheme used. This explains why, e.g., inferring a DCSBM to such a cyclic graph with model selection via an MDL approach [53] distinct from the regularization term used in Eq. (4) results in a split into 22 classes (for a more detailed discussion, see the SM). This “overfitting” behavior is in fact generic and can be observed with many other community detection algorithms, including modularity optimization [55,56] and the map-equation framework [9,57]. The issue is that while the graph structure can be compressed in terms of block structure with relatively small blocks, these blocks are less relevant for the long-term dynamics.

As seen in Fig. 2, for Markov chains with sparse state-transition graphs with long-range path structures, this mismatch between clusters defined via one-step block connectivity ($T = 1$) and clusters capturing the long-term behavior can be quite pronounced. Indeed, the regularized autoinformation for short times is typically optimized by choosing a relatively large number of aggregated states, while the dynamically planted class structure is only found for larger T . This short time behavior of the regularized autoinformation is again mirrored by the MDL-based inference of DCSBMs or the map equation, which both fail to find the dynamically meaningful partition for long timescales for all the graphs shown in Fig. 2: the inference of the DCSBM using the MDL approach in [53] yields around 22 classes in all cases; the map equation provides 7, 10, or 4 aggregation classes for the three scenarios, respectively. While in this case spectral methods such as generalized Perron cluster cluster analysis [13] can resolve the relevant structure, they may fail when intermediate or short timescales are of interest. More in-depth comparisons

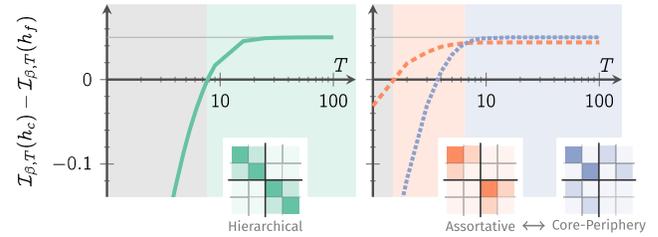


FIG. 3. State aggregations for Markov chains with hierarchical timescales on a stochastic block model. Left: We plot the difference in the regularized autoinformation between a fine-state aggregation h_f into the four planted aggregation classes, and a coarse two-class state aggregation h_c for a hierarchical state-transition graph of a Markov chain (inset). The plot shows that at longer timescales (green shade) $\mathcal{I}_{\beta,T}(h_c) > \mathcal{I}_{\beta,T}(h_f)$ and hence the coarser aggregation is preferred over the fine aggregation, which is preferred at shorter timescales (gray shade). Right: Difference in the regularized autoinformation between the two-class split h_c , describing either a core-periphery (orange) or an assortative (violet) aggregation, and the underlying planted aggregation into four classes (h_f). The four-class partition has a higher autoinformation than the two-class split at short timescales (gray shade). The assortative partition has highest autoinformation for middle range timescales (orange shade), and the core-periphery partition is preferred at longer timescales (blue shade). All graphs consist of 400 nodes and expected average degree $\langle k \rangle = 15$, with $\beta = 0.05$.

can be found in the SM, where we also describe a synthetic model class (a graph ensemble) that displays the behavior observed here. We emphasize that changing the parameters β or T is in general *not equivalent* (see the SM).

Hierarchical aggregation of Markov chains with multiple timescales.—In a hierarchical stochastic block model (see Fig. 3, left), for any given value $\beta > 0$, the finer structure is typically *preferred* at lower values of T where the walker dynamics are confined to the local class. Higher values of T allow the walker to visit larger portions of the network, and coarser partitions gain importance.

Consider now two alternative hierarchical aggregations (core periphery vs assortative) of an initial aggregation into four classes (see Fig. 3, right). While the same four-class structure is preferred at short timescales, the two-class assortative and the core-periphery structures are preferred at medium and longer timescales. Although for two equally sized classes (in terms of entropy), the optimal aggregation is either assortative or disassortative, here, the core and periphery are of different sizes in terms of the probability of the presence of the walker. In particular, the regularization term $\beta H(y_i)$ in Eq. (4) favors the core-periphery split. This effect dominates at large timescales, where the autoinformation converges toward zero. This is intrinsic to what our regularization term in Eq. (4) considers to be a “small” or “simple” model, and other choices of regularization may lead to different results. Similar trade-offs were observed for stochastic block models and DCSBMs in [46,58] when considering core-periphery or assortative structures.

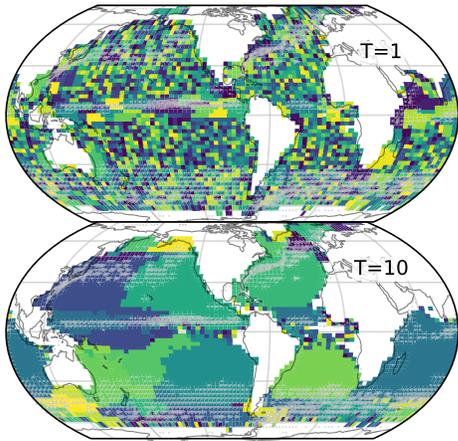


FIG. 4. State aggregation of ocean currents. The above maps compare two partitions induced by aggregating the states of the ocean currents according to the regularized autoinformation for short timescales (top) or longer timescales (bottom) with $\beta = 0.5$. At shorter timescales, a higher number of classes is found. At longer timescales, the aggregation classes reveal well-known features of global ocean dynamics such as the Antarctic Circumpolar Current, subtropical gyres, and, in general, a marked separation of the polar, midlatitude, and equatorial regions. The quiver plot overlay displays the average drifter's velocity. Each time step t corresponds to 16 days.

The system of ocean surface currents.—Let us now showcase how one can use the autoinformation as a tool to analyze dynamical data. The Global Drifters Program [59] tracks drifter buoys on the surface of all oceans. The dynamics of the drifters is a proxy for the global system of surface currents, i.e., water masses moving between different areas of the ocean surfaces.

Using the regularized autoinformation, we identify macro areas that optimally aggregate the drifter dynamics. We find that the temporal dimension of the kinetics strongly influences the outcome. For short timescales, the aggregation classes correspond to small geographic patches of ocean surface that become larger where currents are stronger and steadier, e.g., along the equator (see Fig. 4, top). For timescales closer to the expected time for a drifter to cross an ocean, larger geographic patches are found. These encompass all major ocean gyres (see Fig. 4, bottom) separating equatorial, subtropical and boreal regions. The northern and southern Pacific are subdivided into western and eastern parts that belong to the same large-scale circulation pattern but represent different areas of surface convergence and are located around so-called “garbage patches” [60–62].

Recently, the ocean currents have been clustered in dynamical domains by analyzing a long-term simulation of the barotropic vorticity equation and applying a simple k -means algorithm on the magnitude of the different terms contributing to the vorticity dynamics [63]. This dynamics is only partly comparable to the drifter dynamics as it

involves not just surface currents but an average over all ocean depths. It is nonetheless interesting to compare the outcomes, which share many features (see the SM for these comparisons). However, a key difference is that while the k -means method [63] can lead to geographically disconnected patches, scattered across the globe, our method finds spatially connected classes and is moreover completely data-driven, using only the multiscale dynamical analysis of empirical trajectories.

M. T. S. received funding from the European Unions Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 702410 and the Ministry of Culture and Science (MKW) of the German State of North Rhine-Westphalia (“NRW Rckkehrprogramm”). M. F. was partially funded by Innoviris Grant No. D1.31402.007-F. J. C. D. was partially funded by the Flagship European Research Area Network (FLAG-ERA) Joint Transnational Call FuturICT 2.0. We warmly thank Leto Peel and Eric Deleersnijder for fruitful discussions.

- [1] H. A. Simon and A. Ando, Aggregation of variables in dynamic systems, *Econometrica: J. Econometric Soc.* **29**, 111 (1961).
- [2] B. Moore, Principal component analysis in linear systems: Controllability, observability, and model reduction, *IEEE Trans. Autom. Control* **26**, 17 (1981).
- [3] J.-N. Juang and R. S. Pappa, An eigensystem realization algorithm for modal parameter identification and model reduction, *J. Guid. Control Dyn.* **8**, 620 (1985).
- [4] C. D. Meyer, Stochastic complementation, uncoupling markov chains, and the theory of nearly reducible systems, *SIAM Rev.* **31**, 240 (1989).
- [5] J. P. Crutchfield and K. Young, Inferring Statistical Complexity, *Phys. Rev. Lett.* **63**, 105 (1989).
- [6] R. R. Coifman and S. Lafon, Diffusion maps, *Appl. Comput. Harmon. Anal.* **21**, 5 (2006).
- [7] W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen, The multiscale coarse-graining method. i. a rigorous bridge between atomistic and coarse-grained models, *J. Chem. Phys.* **128**, 244114 (2008).
- [8] W. H. A. Schilders, H. A. Van der Vorst, and J. Rommes, *Model Order Reduction: Theory, Research Aspects and Applications* (Springer, New York, 2008), Vol. 13, <https://doi.org/10.1007/978-3-540-78841-6>.
- [9] M. Rosvall and C. T. Bergstrom, Maps of random walks on complex networks reveal community structure, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 1118 (2008).
- [10] J.-C. Delvenne, S. N. Yaliraki, and M. Barahona, Stability of graph communities across time scales, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 12755 (2010).
- [11] J. Nathan Kutz, S. L. Brunton, B. W. Brunton, and Joshua L. Proctor, *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems* (SIAM, Philadelphia, 2016), Vol. 149, <https://doi.org/10.1137/1.9781611974508>.

- [12] P. Benner, A. Cohen, M. Ohlberger, and K. Willcox, *Model Reduction and Approximation: Theory and Algorithms* (SIAM, Philadelphia, 2017), Vol. 15, <https://doi.org/10.1137/1.9781611974829>.
- [13] K. Fackeldey, A. Sikorski, and M. Weber, Spectral clustering for non-reversible markov chains, *Comput. Appl. Math.* **37**, 6376 (2018).
- [14] M. Coderch, A. Willsky, S. Sastry, and D. Castanon, Hierarchical aggregation of linear systems with multiple time scales, *IEEE Trans. Autom. Control* **28**, 1017 (1983).
- [15] N. Monshizadeh, H. L. Trentelman, and M. Kanat Camlibel, Projection-based model reduction of multi-agent systems using graph partitions, *IEEE Trans. Control Network Systems* **1**, 145 (2014).
- [16] M. T. Schaub, N. O'Clery, Y. N. Billeh, J.-C. Delvenne, R. Lambiotte, and M. Barahona, Graph partitions and cluster synchronization in networks of oscillators, *Chaos* **26**, 094821 (2016).
- [17] N. Tishby, F. C. Pereira, and W. Bialek, The information bottleneck method, [arXiv:physics/0004057](https://arxiv.org/abs/physics/0004057).
- [18] C. Rohilla Shalizi and J. P. Crutchfield, Computational mechanics: Pattern and prediction, structure and simplicity, *J. Stat. Phys.* **104**, 817 (2001).
- [19] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevLett.127.078301> for additional mathematical proofs, information on introduced examples and deeper comparison with related works, which includes Refs. [20–44].
- [20] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, New York, 2001), <https://doi.org/10.1002/0471200611>.
- [21] M. E. J. Newman and T. P. Peixoto, Generalized Communities in Networks, *Phys. Rev. Lett.* **115**, 088701 (2015).
- [22] M. T. Schaub, J.-C. Delvenne, R. Lambiotte, and M. Barahona, Multiscale dynamical embeddings of complex networks, *Phys. Rev. E* **99**, 062308 (2019).
- [23] M. E. J. Newman, Assortative Mixing in Networks, *Phys. Rev. Lett.* **89**, 208701 (2002).
- [24] J.-C. Delvenne, M. T. Schaub, S. N. Yaliraki, and M. Barahona, The stability of a graph partition: A dynamics-based framework for community detection, in *Dynamics On and Of Complex Networks, Volume 2* (Springer, New York, 2013), pp. 221–242, https://doi.org/10.1007/978-1-4614-6729-8_11.
- [25] L. Berline, A.-M. Rammou, A. Doglioli, A. Molcard, and A. Petrenko, A connectivity-based eco-regionalization method of the mediterranean sea, *PLoS One* **9**, e111978 (2014).
- [26] V. Rossi, E. Ser-Giacomi, C. López, and E. Hernández-García, Hydrodynamic provinces and oceanic connectivity from a transport network help designing marine reserves, *Geophys. Res. Lett.* **41**, 2883 (2014).
- [27] C. J. Thomas, J. Lambrechts, E. Wolanski, V. A. Traag, V. D. Blondel, E. Deleersnijder, and E. Hanert, Numerical modelling and graph theory tools to study ecological connectivity in the great barrier reef, *Ecol. Model.* **272**, 160 (2014).
- [28] C. Wunsch and D. Stammer, Satellite altimetry, the marine geoid, and the oceanic general circulation, *Annu. Rev. Earth Planetary Sci.* **26**, 219 (1998).
- [29] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, Complex networks in climate dynamics, *Eur. Phys. J. Special Topics* **174**, 157 (2009).
- [30] N. Molkenhain, K. Rehfeld, N. Marwan, and J. Kurths, Networks from flows—from dynamics to topology, *Sci. Rep.* **4**, 4119 (2014).
- [31] L. Tupikina, N. Molkenhain, C. Lpez, E. Hernández-Garca, N. Marwan, and J. Kurths, Correlation networks from flows. the case of forced and time-dependent advection-diffusion dynamics, *PLoS One* **11**, e0153703 (2016).
- [32] J. P. Crutchfield and D. P. Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, *Chaos* **13**, 25 (2003).
- [33] D. Kelly, M. Dillingham, A. Hudson, and K. Wiesner, A new method for inferring hidden markov models from noisy time sequences, *PLoS One* **7**, e29703 (2012).
- [34] N. Masuda, M. A. Porter, and R. Lambiotte, Random walks and diffusion on networks, *Phys. Rep.* **716–717**, 1 (2017).
- [35] T. P. Peixoto and M. Rosvall, Modelling sequences and temporal networks with dynamic community structures, *Nat. Commun.* **8**, 582 (2017).
- [36] M. E. J. Newman and M. Girvan, Finding and evaluating community structure in networks, *Phys. Rev. E* **69**, 026113 (2004).
- [37] M. T. Schaub, Unraveling complex networks under the prism of dynamical processes: relations between structure and dynamics, Ph.D. thesis, Imperial College London, 2014, <https://doi.org/10.25560/38446>.
- [38] M. E. J. Newman, Equivalence between modularity optimization and maximum likelihood methods for community detection, *Phys. Rev. E* **94**, 052315 (2016).
- [39] P. Grindrod, Range-dependent random graphs and their application to modeling large small-world proteome datasets, *Phys. Rev. E* **66**, 066702 (2002).
- [40] M. Ángeles Serrano, D. Krioukov, and Marián Boguñá, Self-Similarity of Complex Networks and Hidden Metric Spaces, *Phys. Rev. Lett.* **100**, 078701 (2008).
- [41] P. Buchholz, Exact and ordinary lumpability in finite markov chains, *J. Appl. Probab.* **31**, 59 (1994).
- [42] C. Godsil and G. F. Royle, *Algebraic Graph Theory* (Springer Science & Business Media, New York, 2013), Vol. 207.
- [43] J. Paul Tian and D. Kannan, Lumpability and commutativity of markov processes, *Stoch. Anal. Appl.* **24**, 685 (2006).
- [44] N. O'Clery, Y. Yuan, G.-B. Stan, and M. Barahona, Observability and coarse graining of consensus dynamics through the external equitable partition, *Phys. Rev. E* **88**, 042805 (2013).
- [45] A. Dasgupta, J. E. Hopcroft, and F. McSherry, Spectral analysis of random graphs with skewed degree distributions, in *45th Annual IEEE Symposium on Foundations of Computer Science* (IEEE, New York, 2004), pp. 602–610, <https://doi.org/10.1109/FOCS.2004.61>.
- [46] B. Karrer and M. E. J. Newman, Stochastic blockmodels and community structure in networks, *Phys. Rev. E* **83**, 016107 (2011).
- [47] S. Fortunato, Community detection in graphs, *Phys. Rep.* **486**, 75 (2010).

- [48] S. Fortunato and D. Hric, Community detection in networks: A user guide, *Phys. Rep.* **659**, 1 (2016).
- [49] M. T. Schaub, J.-C. Delvenne, M. Rosvall, and R. Lambiotte, The many facets of community detection in complex networks, *Appl. Network Sci.* **2**, 4 (2017).
- [50] M. Faccin, M. T. Schaub, and J.-C. Delvenne, Entrograms and coarse graining of dynamics on complex networks, *J. Complex Netw.* **6**, 661 (2018).
- [51] P. D. Grünwald and A. Grunwald, *The Minimum Description Length Principle* (MIT Press, Cambridge, MA, 2007), <https://doi.org/10.7551/mitpress/4643.001.0001>.
- [52] H. Akaike, A new look at the statistical model identification, *IEEE Trans. Autom. Control* **19**, 716 (1974).
- [53] T. P. Peixoto, Parsimonious Module Inference in Large Networks, *Phys. Rev. Lett.* **110**, 148701 (2013).
- [54] Code repository available at <https://maurofaccin.github.io/aisa>.
- [55] A. Clauset, M. E. J. Newman, and C. Moore, Finding community structure in very large networks, *Phys. Rev. E* **70**, 066111 (2004).
- [56] M. T. Schaub, J.-C. Delvenne, S. N. Yaliraki, and M. Barahona, Markov dynamics as a zooming lens for multiscale community detection: Non clique-like communities and the field-of-view limit, *PLoS One* **7**, e32210 (2012).
- [57] M. T. Schaub, R. Lambiotte, and M. Barahona, Encoding dynamics for multiscale community detection: Markov time sweeping for the map equation, *Phys. Rev. E* **86**, 026112 (2012).
- [58] L. Peel, D. B. Larremore, and A. Clauset, The ground truth about metadata and community detection in networks, *Sci. Adv.* **3**, e1602548 (2017).
- [59] Global Drifter Program: <http://www.aoml.noaa.gov/phod/gdp/index.php>.
- [60] L. C. Young, C. Vanderlip, D. C. Duffy, V. Afanasyev, and S. A. Shaffer, Bringing home the trash: Do colony-based differences in foraging distribution lead to increased plastic ingestion in laysan albatrosses? *PLoS One* **4**, e7623 (2009).
- [61] E. A. Howell, S. J. Bograd, C. Morishige, M. P. Seki, and J. J. Polovina, On north pacific circulation and associated marine debris concentration, *Mar. Pollut. Bull.* **65**, 16 (2012), at-sea Detection of Derelict Fishing Gear.
- [62] L. Lebreton, B. Slat, F. Ferrari, B. Sainte-Rose, J. Aitken, R. Marthouse, S. Hajbane, S. Cunsolo, A. Schwarz, A. Levivier, K. Noble, P. Debeljak, H. Maral, R. Schoeneich-Argent, R. Brambini, and J. Reisser, Evidence that the great pacific garbage patch is rapidly accumulating plastic, *Sci. Rep.* **8**, 4666 (2018).
- [63] M. Sonnewald, C. Wunsch, and P. Heimbach, Unsupervised learning reveals geography of global ocean dynamical regions, *Earth and Space Sci.* **6**, 784 (2019).