Variational kinetic clustering of complex networks

Vladimir Koskin; Adam Kells ; Joe Clayton; Alexander K. Hartmann ; Alessia Annibale ; Edina Rosta

Check for updates

CrossMark

View Online

Export Citation

Articles You May Be Interested In

Variational kinetic clustering of complex networks

View Online   Export Citation   CrossMark

Vladimir Koskin,[1,2] Adam Kells,[1] (ID) Joe Clayton,[2] Alexander K. Hartmann,[3] (ID) Alessia Annibale,[4,a)] (ID)
and Edina Rosta[2,a)] (ID)

AFFILIATIONS

[1] Department of Chemistry, King's College London, SE1 1DB London, United Kingdom
[2] Department of Physics and Astronomy, University College London, WC1E 6BT London, United Kingdom
[3] Institute of Physics, University of Oldenburg, Oldenburg, Germany
[4] Department of Mathematics, King's College London, SE11 6NJ London, United Kingdom

[a)] Authors to whom correspondence should be addressed: alessia.annibale@kcl.ac.uk and e.rosta@ucl.ac.uk

ABSTRACT

Efficiently identifying the most important communities and key transition nodes in weighted and unweighted networks is a prevalent problem in a wide range of disciplines. Here, we focus on the optimal clustering using variational kinetic parameters, linked to Markov processes defined on the underlying networks, namely, the slowest relaxation time and the Kemeny constant. We derive novel relations in terms of mean first passage times for optimizing clustering via the Kemeny constant and show that the optimal clustering boundaries have equal round-trip times to the clusters they separate. We also propose an efficient method that first projects the network nodes onto a 1D reaction coordinate and subsequently performs a variational boundary search using a parallel tempering algorithm, where the variational kinetic parameters act as an energy function to be extremized. We find that maximization of the Kemeny constant is effective in detecting communities, while the slowest relaxation time allows for detection of transition nodes. We demonstrate the validity of our method on several test systems, including synthetic networks generated from the stochastic block model and real world networks (Santa Fe Institute collaboration network, a network of co-purchased political books, and a street network of multiple cities in Luxembourg). Our approach is compared with existing clustering algorithms based on modularity and the robust Perron cluster analysis, and the identified transition nodes are compared with different notions of node centrality.

I. INTRODUCTION

Networks[1–8] (or graphs) are a powerful tool to model systems of many variables, with complex patterns of interactions. Examples are found in virtually every field, ranging from biology[9–14] to finance,[15–18] social sciences[19–25] to law,[26,27] and literature.[28,29]

Coarse-grained representations of large complex networks are often used to gain intuition, which usually require the identification of clusters (or communities).[30–34] Algorithms for cluster detection in unweighted networks typically rely on topological features of the network, i.e., they depend only on the usually static connection patterns and not on the processes that are taking place in the networks. Exact approaches consider all the different partitions into subnetworks and optimize suitable quantities, such as

the so-called modularity.[35,36] However, these scale poorly with the size of the system: when clustering a network with N nodes into a coarse-grained one with m nodes, there are m^N possible partitions. Many clustering problems, therefore, require a search in an exponentially increasing space with the system size, leading to non-deterministic polynomial-time (NP) hard problems.[37,38] Local, incremental, and/or deterministic search algorithms, therefore, may be unable to identify the global optimum, such as most versions of the popular k-means clustering,[39] and thus more advanced global search algorithms are needed.

Such complexity also arises when performing kinetic clusterings of Markov processes. These are memoryless dynamical processes evolving on a finite set of states, which can be seen as weighted networks, where the nodes represent the Markov states

19 September 2023 21:32:22

and the edge weights are given by the transition rates. In molecular sciences, the states of such kinetic networks typically represent selected conformational states of the system, e.g., local minima and saddle points of an energy landscape, and transition rates between them are defined in terms of their energy barriers.[30,40–45] More recently, such dynamical networks are derived from molecular dynamics (MD) simulations by defining appropriate Markov State Models (MSMs).[46–50] Markov processes have recently become a prominent tool for modeling and interpreting large simulation data-sets of complex kinetic systems in many other domains of academia and industry.[51–53]

Different coarse-graining techniques can be used to reduce the dimensionality of MSMs. One option is to group microstates of the system (nodes) together into macrostates (clusters). These clusters are normally taken as the metastable states of the system, i.e., sets of microstates between which the system is slow to move. Hence, many algorithms for kinetic clustering are related to maximizing the metastability of the clusters.[54–65] Alternative approaches can be used, e.g., spectral matching.[66,67]

In this work, we focus on optimal coarse-graining based on lumping microstates into clusters. Such clusters are then used to define coarse-grained versions of the original system, where the nodes (or states) are the clusters and the original kinetics is approximated by effective transitions between clusters.[68] This leads to a transition rate matrix in coarse-grained space, whose eigenvalues and eigenvectors will differ from those of the original transition matrix, thus resulting in different kinetics. An alternative to maximizing the metastability of the clusters is to maximize the timescales of the eigenprocesses in the clustered system, aiming to variationally preserve the slowest relaxation times.

The Kemeny constant[69] describes the expected time to reach any target microstate, averaged over all microstates. It can be expressed as the sum of the timescales of all the eigenprocesses, defined by the eigenvalues of the transition rates matrix. It was shown that a coarse-grained MSM satisfies a variational principle with respect to the Kemeny constant.[70] Hence, the Kemeny constant is an obvious candidate as a variational parameter to be maximized when performing clustering, as conjectured in recent literature.[71–73]

In addition to clusters, the notion of transition states (TS) is of particular importance in the context of MSMs: these are bottleneck states that the system passes through while moving between the metastable clusters. A related notion in network science is the one of node centrality. Different measures of node centrality have been introduced to characterize nodes linking between two or more clusters, e.g., betweeness[74] and closeness[75] centrality; however, these are based on topological features of the network and disregard kinetics information, and hence their use to automatically identifying transition states in kinetic and, more in general, weighted networks, remains unclear. In earlier work, it was demonstrated that the slowest timescale in kinetic networks can be used as a variational parameter for finding transition states effectively;[76,77] however, the algorithm developed in Ref. 76 was inefficient when scaled to high-dimensional systems.

In this work, building on the framework developed in Refs. 70 and 76, we propose a new clustering method based on the optimization of the Kemeny constant of the coarse-grained system, which satisfies a variational principle with respect to the original dynamics and is effective in finding metastable clusters. We derive properties of the optimal boundary positions in terms of mean first passage times (MFPTs). We find that in the optimal clustering of large complex networks each node belongs to its nearest cluster measured via round-trip times distance.

Our novel implementation aims at accelerating the search for optimal clustering using *parallel tempering*,[78–80] which is an efficient global and stochastic optimization based on interpreting the target function to optimize as a physical energy and coupling the optimization process to several artificial heat baths in parallel.[81,82] Our algorithm can similarly be applied to optimize other objective functions, including the slowest timescale used in Ref. 76 to detect transition states. In this way, we provide a computationally efficient way to automatically detect communities as well as transition clusters in complex networks, based on kinetic properties of network processes.

In Sec. II, we lay down the theoretical framework of the present work by describing the concepts of network clustering for kinetic and unweighted networks. In Sec. III, we present new analytical expressions for the derivative of the Kemeny constant with respect to the cluster boundary positions for arbitrary number of states and show that the optimal boundaries correspond to equal round-trip times between the clusters they separate. We also present a clustering algorithm, which we name the "parallel tempering variational clustering" (PTVC), for optimizing clusters using arbitrary objective functions on complex networks, based on the parallel tempering method used in statistical physics. Finally, in Sec. III D, we compare clustering resulting from optimizing different objective functions, i.e., modularity, stability, and Kemeny constant, as well as using the Perron cluster analysis on synthetic and real-world networks.

II. THEORY

A. Clustering in complex networks

When working with large or complex networks, one often encounters problems with visualization and interpretation of observations due to the absence of a low dimensional space to project the graph onto. Often, it is helpful to convert a large network into a smaller one, while preserving the features of interest from the original network. This conversion is known as network clustering,[83] coarse-graining,[84] or graph partitioning[85] and is equivalent to assigning each node of the graph to a community (cluster).[86]

The clustering of an N-node network into m clusters can be formally defined by a rectangular $N \times m$ crisp assignment matrix $\mathbf{S}$, where each entry $S_{iI} \in \{0, 1\}$ indicates whether (1) or not (0) node $i \in \{1, \ldots, N\}$ belongs to cluster $I \in \{1, \ldots, m\}$, satisfying $\sum_{I=1}^{m} S_{iI} = 1$.

Choosing $\mathbf{S}$ optimally usually entails maximizing or minimizing a certain objective function, which depends on the partitioning itself. It has been widely accepted that the objective function to use depends on the system modeled by the network and the exact task at hand; hence, many metrics have been introduced to quantify clustering quality,[87,88] which can be separated into two classes: topology-based and kinetics-based. Kinetic clustering uses kinetic properties of Markov processes, and it is widely used to cluster kinetic networks, whereas topology-based clustering relies on structural properties of the network and it is more common with unweighted networks.

1. Kinetic networks

Any finite Markov process can be modeled as a weighted directed network, also called kinetic network, where nodes represent the discrete states of the Markov process and edge weights represent the transition rates between different states. The state occupancy probabilities evolve according to the master equation, which relates their rate of change to the difference in the probability flux in and out of the states

$$\frac{dp_j(t)}{dt} = \sum_{i(\neq j)} \left[K_{ji} p_i(t) - K_{ij} p_j(t) \right], \tag{1}$$

where $p_j(t)$ is the occupancy probability of state j and K_{ji} is the transition rate from state i to state j, with $K_{jj} = -\sum_{i(\neq j)} K_{ij}$ being the rate of exits from state j. Inserting this definition into Eq. (1) one arrives at the matrix form

$$\frac{d\mathbf{p}}{dt} = \mathbf{K}\mathbf{p}. \tag{2}$$

All of the kinetic information of the system is encoded in the eigenvalues λ_n of the rate matrix $\mathbf{K}$, where $n \in \{1, \ldots, N\}$ for a system with N states, and their corresponding (right and left) eigenvectors.

If the Markov process is irreducible, one of the eigenvalues will be zero and the rest will all have a negative real part. We will focus on systems satisfying detailed balance, where the eigenvalues are guaranteed to be real and can be ordered as follows:

$$0 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_N. \tag{3}$$

From Eq. (2), it can be seen that the stationary (equilibrium) distribution $\boldsymbol{\pi}$ is given by the right eigenvector corresponding to the zero eigenvalue λ_1. The other eigenvalues are related to the timescales τ_n with which the rate matrix moves probability density between the oppositely signed regions of the corresponding eigenvector[89]

$$\tau_n = -\frac{1}{\lambda_n} \quad n \in \{2, \ldots, N\}. \tag{4}$$

When the Markov process evolves at discrete times $t = n\tau$ ($n \in \mathbb{N}$), the evolution of the state occupancy probabilities is formulated in terms of the Markov chain

$$\mathbf{p}(n) = \mathbf{M}^n(\tau)\mathbf{p}(0), \tag{5}$$

where $\mathbf{p}(n)$ is the probability at time step n and $\mathbf{M}(\tau)$ is the matrix of transition probabilities between pairs of states over the lagtime τ. Trivially, Eq. (5) is equivalent to Eq. (2) when identifying $\mathbf{M}(\tau) = e^{\mathbf{K}\tau}$.

2. Unweighted networks

Unweighted networks are defined by an adjacency matrix $\mathbf{A}$, where each entry $A_{ij} \in \{0, 1\}$ determines whether (1) or not (0) an edge is present between nodes i and j, for $i, j \in \{1, \ldots, N\}$. In general, a network may be directed. For undirected networks, the adjacency matrix is symmetric, i.e., $A_{ij} = A_{ji} \, \forall \, i, j$.

It is possible to define a diffusion process on an unweighted network to transforms it into a kinetic network. This allows for the application of kinetic clustering methods. For diffusion processes on unweighted networks, the rate matrix $\mathbf{K}$ is given by the random walk

normalized Laplacian matrix $\mathbf{L}$[90] via $\mathbf{K} \equiv -\mathbf{L}$. The matrix $\mathbf{L}$ is defined by

$$\mathbf{L} \equiv \mathbf{I} - \mathbf{A}\mathbf{\Delta}^{-1}, \tag{6}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{\Delta}$ is a diagonal matrix with elements $\Delta_{ij} = d_i \delta_{ij}$, where $d_i = \sum_{j=1}^{N} A_{ij}$ is the degree of node i.

Alternatively, one can define a discrete-time random walk on the network links, by means of the transition matrix

$$\mathbf{M}(\tau) = \left(\mathbf{A}\mathbf{\Delta}^{-1} \right)^\tau \tag{7}$$

for an arbitrary integer lagtime τ.

Since the Laplacian [or, equivalently, the transition matrix given in Eq. (7)] is fully determined by the adjacency matrix, kinetic clustering based on diffusion or random walks is to make contact with topology-based clustering, relying purely on network structure.

B. Modularity

A popular metric to assess the quality of a cluster assignment $\mathbf{S}$ in unweighted networks is the network modularity.[35,91] For undirected networks, the modularity Q is defined as

$$Q(\mathbf{S}) = \frac{1}{N\bar{d}} \sum_{J=1}^{m} \sum_{i,j} S_{iJ} \left(A_{ij} - \frac{d_i d_j}{N\bar{d}} \right) S_{jJ}, \tag{8}$$

where

$$\bar{d} = N^{-1} \sum_{i=1}^{N} d_i$$

is the average degree of the network.

The most widely used methods to cluster unweighted networks aim at finding the assignment matrix $\mathbf{S}$ that maximizes the network modularity Q, and several greedy optimization algorithms have been developed to handle large networks with over 100×10^6 nodes.[92,93]

However, the modularity suffers from a well-known resolution limit[94] that prevents it from resolving small communities, and it has been recently criticized as a metrics for information recovery in a network.[95] Moreover, modularity is only based on structural properties of networks; hence, it does not take into account the kinetic process that a network may support. However, it has been shown to be equivalent to a kinetic clustering quality metric named *stability*,[96] when applied to diffusion processes on unweighted networks.

C. Kinetic clustering

Optimizing kinetic clustering by lumping together microstates and defining the best cluster assignment $\mathbf{S}$ can be achieved via appropriate objective functions, which are designed to enforce desired kinetic properties in the coarse-grained system. Usually, they depend on the spectral properties of the Markov process (or Markov chain) defined in the original state space. A prominent example is the popular Perron cluster–cluster analysis (PCCA),[56] and its improved version PCCA+, introduced by Deuflhardt and Weber.[62,97–99] An alternative approach is to optimize objective functions that encode spectral properties of the *coarse-grained* dynamics.[83,100] An advantage of this method is that certain spectral quantities of the coarse-grained dynamics satisfy a variational principle with respect to the

original dynamics[70,77] and thus provide a framework for performing a clustering that is variationally optimal.

Below, we review these methods, and, building on the framework presented in Ref. 70, we propose a new clustering method based on the optimization of the Kemeny constant of the coarse-grained system.

1. Coarse-grained Markovian processes

Coarse-graining a Markov process formally corresponds to projecting the Markovian dynamics onto a lower dimensional space, which generally introduces memory effects and consequently loss of the Markovian property. Hence, a question arises as to what the best Markovian approximation is of the resulting non-Markovian dynamics.

For a given clustering $\mathbf{S}$, it has been shown that projections preserving detailed balance are those for which the Laplace transformed equilibrium correlation matrix

$$\hat{\mathbf{C}}(s) = \int_0^\infty \mathbf{C}(\tau)e^{-s\tau}d\tau \qquad (9)$$

in the original and clustered systems satisfy the following relation:[77]

$$\hat{C}_{JI}^{CG}(s) = \sum_{ij=1}^N \hat{C}_{ji}(s)S_{iI}S_{jJ}. \qquad (10)$$

The entries of the correlation matrix $\mathbf{C}(\tau)$ are the equilibrium connected correlation functions $C_{ij}(\tau)$ of the state occupancy for every pair of states (i,j), in the original Markovian system, defined as

$$C_{ji}(\tau) = M_{ji}(\tau)\pi_i - \pi_i\pi_j, \qquad (11)$$

and $C_{JI}^{CG}(\tau)$ is the equilibrium connected correlation function of the cluster occupancy for the pair of clusters (I, J) in the coarse-grained system.

The definition of an effective Markovian dynamics in the lower dimensional space requires approximations in which Eq. (10) is only satisfied in specific limits. Different definitions have been considered for an effectively Markovian coarse-grained dynamics, which correspond to different requirements on the correlation function of the coarse-grained system. These include the local equilibrium (LE) and the recently introduced Hummer–Szabo (HS)[100] method.

The local equilibrium (LE) reduction method is a popular choice, with numerically stable implementations also available via graph transformation.[101] It consists in equating the correlation matrices at a specific finite value of τ,

$$C_{JI}^{LE}(\tau) = \sum_{ij=1}^N C_{ji}(\tau)S_{iI}S_{jJ}. \qquad (12)$$

Note that this means that the correlation matrices will depend on the lagtime τ and that, depending how the correlation is used to obtain a clustering $\mathbf{S}$, the optimum result may depend on τ. The equation reads, in matrix form,

$$\mathbf{C}^{LE}(\tau) = \mathbf{S}^T\mathbf{C}(\tau)\mathbf{S}. \qquad (13)$$

Hence, it corresponds to the $s \to \infty$ limit of Eq. (10). Equation (13), by using Eq. (11) and a corresponding one for $\mathbf{C}^{LE}(\tau)$, provides

a relation for the transition matrix $\mathbf{M}^{LE}(\tau)$ of the coarse-grained dynamics

$$\mathbf{M}^{LE}(\tau)\mathbf{D}_\Pi = \mathbf{S}^T\mathbf{M}(\tau)\mathbf{D}_\pi\mathbf{S}, \qquad (14)$$

where $\mathbf{D}_\pi$ and $\mathbf{D}_\Pi$ are diagonal matrices with the stationary distributions of the original and coarse-grained systems on the diagonal, respectively. Thus, $\mathbf{D}_\Pi = \mathbf{S}^T\mathbf{D}_\pi$.

On the other hand, the Hummer–Szabo (HS) method equates the integral of the correlation matrices over all lagtimes τ, and thus it corresponds to the $s \to 0$ limit of Eq. (10),

$$\int_0^\infty C_{JI}^{HS}(\tau)d\tau = \sum_{i\in I}\sum_{j\in J}\int_0^\infty C_{ji}(\tau)d\tau. \qquad (15)$$

In the matrix form, we have

$$\hat{\mathbf{C}}^{HS} = \mathbf{S}^T\hat{\mathbf{C}}\mathbf{S}, \qquad (16)$$

where we have used the short-hand notation $\hat{\mathbf{C}} = \hat{\mathbf{C}}(0)$ to denote the time-integrated connected correlation function.

Equation (16) provides a different definition of the coarse-grained dynamics with respect to Eq. (13) for the same clustering $\mathbf{S}$, and it leads to a formulation of the coarse-grained dynamics, that in continuous time is given in terms of the rate matrix[100]

$$\mathbf{K}^{HS} = \mathbf{S}^T\boldsymbol{\pi}\mathbf{1}_m^T - \mathbf{D}_\Pi(\mathbf{S}^T(\boldsymbol{\pi}\mathbf{1}_N - \mathbf{K})^{-1}\mathbf{D}_\pi\mathbf{S})^{-1}, \qquad (17)$$

where $\mathbf{1}_m$ is a vector of length m with entries equal to one. The HS method is known to provide numerically equivalent results to the LE method at long lagtimes in many model systems.[76]

In the present work, we will use the HS definition of the coarse-grained dynamics, as it does not require the choice of a specific lag-time, it guarantees that MFPTs in the clustered dynamics match the weighted MFPTs of the microscopic dynamics[70] and it has been shown to replicate the dynamics of the original system more closely than the LE method.[100]

Below, we review different methods to identify the cluster assignment $\mathbf{S}$.

2. Stability

The approach proposed by Barahona et al.[96,102] to perform a kinetic clustering is to maximize a quantity called stability, defined as the sum of the connected autocovariances of the clusters J at a specific lagtime τ,

$$\Omega^S(\tau) \equiv \sum_{J=1}^m C_{JJ}^{LE}(\tau) = \mathrm{Tr}(\mathbf{C}^{LE}(\tau)), \qquad (18)$$

where $\mathbf{C}^{LE}$ is calculated using the LE method as in Eq. (13). The authors have shown that, for a random walk on an unweighted non-directed network, such that $\mathbf{M}$ is given in Eq. (7) and $\pi_i = d_i/N\bar{d}$, this quantity is equivalent to the network modularity when the time parameter τ is set to one

$$\Omega^S(1) = \sum_{J=1}^m \sum_{i,j\in J}\left(\frac{A_{ji}}{d_i}\frac{d_i}{N\bar{d}} - \frac{d_i}{N\bar{d}}\frac{d_j}{N\bar{d}}\right) = Q. \qquad (19)$$

The authors of the method have proposed to optimize the quantity defined in Eq. (18) over the assignment matrix $\mathbf{S}$ for varying values of τ.

Since the optimal number of clusters decreases monotonically when increasing τ, this approach introduces a dependence of the optimal clustering on the parameter τ, which controls the resolution of the clustering. This allows for the identification of multiple clusterings, with finer or coarser structure, for a single network, but it increases the overall computational cost by requiring to optimize the clustering for each value of τ to identify the relevant number of clusters.[96]

3. Perron clustering

The Perron cluster–cluster analysis (PCCA) is a method for identifying communities or clusters in *nearly uncoupled* Markov chains.[56] It exploits the fact that in uncoupled Markov chains, characterized by a block-diagonal transition matrix, the entries of the left eigenvectors are constant on each cluster. This allows us to regroup nodes following the sign structure of the left eigenvectors. The assumption is that this sign structure remains stable under small perturbations, i.e., the transition matrix remains block-diagonal dominant, so that a similar regrouping is possible in nearly uncoupled Markov chains.

The original method has been shown to suffer from lack of robustness, and the PCCA+ has been developed to use in practical applications.[62] The main difference from PCCA is the introduction of a soft (fuzzy) assignment matrix $\hat{\mathbf{S}}$, where each node i is assigned a clustering vector of length m (number of clusters) satisfying the conditions of positivity and partition of unity

$$\hat{S}_{iI} \in [0,1] \quad \forall \; i, I, \tag{20}$$

$$\sum_{I=1}^{m} \hat{S}_{iI} = 1 \quad \forall \; i, \tag{21}$$

hence interpretable as the probability of node i to belong to each cluster.[99]

The main assumption of PCCA+ is that the assignment matrix $\hat{\mathbf{S}}$ can be related to the eigenvectors of the transition matrix of the original system, by a transformation matrix $\mathbf{T} \in \mathbb{R}^{m \times m}$ via

$$\hat{\mathbf{S}} = \mathbf{XT}, \tag{22}$$

where $\mathbf{X} = [\boldsymbol{\Psi}_1^L, \dots, \boldsymbol{\Psi}_m^L] \in \mathbb{R}^{n \times m}$ and $\boldsymbol{\Psi}_k^L$ is the left eigenvector of the transition matrix associated with the kth largest eigenvalue, normalized in such a way that $\mathbf{X}^T \mathbf{D}_\pi \mathbf{X} = \mathbf{I}$. The clustering problem then consists in finding the matrix $\mathbf{T}$ that optimizes an objective function under the constraints defined in Eqs. (22), (20), and (21). Optimizing $\mathbf{T}$, rather than $\mathbf{S}$, leads to a reduction of the number of variables to be optimized from $(n \times m)$ to (m^2).

Weber originally proposed the objective function, here denoted as Ω^{PCCA+W},

$$\Omega^{PCCA+W} \equiv \mathrm{Tr}\left(\hat{\mathbf{S}}^T \mathbf{M}(\tau) \mathbf{D}_\pi \hat{\mathbf{S}} \mathbf{D}_\Pi^{-1}\right), \tag{23}$$

which can be written, for crisp assignment $\mathbf{S}$, as the trace of the coarse-grained transition matrix obtained using the LE method defined in Eq. (14).[97]

However, it was noted in Ref. 98 that the interpretation of Ω^{PCCA+W} as the trace of a coarse-grained transition matrix is not valid for soft assignment matrices; so, Roblitz proposed a new objective function[99]

$$\Omega^{PCCA+R} \equiv \mathrm{Tr}\left(\hat{\mathbf{S}}^T \mathbf{D}_\pi \hat{\mathbf{S}} \mathbf{D}_\Pi^{-1}\right) = \mathrm{Tr}\left(\mathbf{T}^T \mathbf{T} \mathbf{D}_\Pi^{-1}\right), \tag{24}$$

where the second equality uses Eq. (22) and normalization of the eigenvectors matrix $\mathbf{X}$. Given the property of $\hat{\mathbf{S}}$ stated in Eq. (21), the matrix $\hat{\mathbf{S}}^T \mathbf{D}_\pi \hat{\mathbf{S}} \mathbf{D}_\Pi^{-1}$ is stochastic; hence, its trace is upper bounded by its dimension m, with the value m being attained by *any* crisp clustering $\mathbf{S}$. Thus, maximizing Ω^{PCCA+R} is equivalent to making the clustering $\hat{\mathbf{S}}$ as crisp as possible.

As the maximization of Eq. (24) over the entries of $\mathbf{T}$ is subject to the constraints Eqs. (22), (20), and (21), the optimal $\mathbf{T}$ will depend on the eigenvectors of the transition matrix $\mathbf{M}(\tau)$, through $\mathbf{X}$. Since in a truly Markovian system the eigenvectors of the transition matrix are independent of the lagtime τ, there is, in principle, no need to optimize over different values of τ, as in the stability method. However, it has to be noted that molecular simulations often exhibit non-Markovian behavior; hence, a dependence on the lagtime is expected, in practice, when Markov matrices are constructed using simulation data.

In most applications, the PCCA+ objective function defined in Eq. (24) is used. In our implementation of the PCCA+ method, we made the same choice and used the efficient Schur decomposition to compute $\mathbf{X}$, following Weber et al.[103,104]

4. Slowest timescale

Recently, a variational kinetic clustering has been proposed in Ref. 76 that uses crisp assignment matrices and aims to maximize the second largest eigenvalue of the clustered dynamics. It was shown that this method is effective in identifying transition states, alongside to key metastable states.

This approach has been shown to be variationally optimal in Ref. 77, where it was proven that the slowest timescale of the clustered system is always smaller or equal to the one of the original network

$$\tau_2^{orig} \geq \tau_2^{CG}, \tag{25}$$

regardless of the protocol (LE or HS) used to coarse-grain the dynamics.

However, the method has only been tested on systems that contain either a small number of metastable states or a spectral gap between the slowest and second slowest timescales. The performance of this method on systems with several metastable states, where $\tau_n \approx \tau_2$ for $n > 2$, remains to be investigated. The existence of multiple slow processes of similar timescale may require a variational parameter that incorporates multiple timescales. An interesting generalization of the slowest timescale would be considering an objective function that is the sum of the largest relaxation times

$$\Omega^{Sum(h)} \equiv \sum_{i=2}^{h+1} \tau_i^{CG}, \tag{26}$$

where h is the number of timescales considered.

5. Kemeny constant

In Markovian systems, the sum of all timescales is known as the Kemeny constant, which is also equivalent with the weighted sum of all mean first passage times (MFPTs) t_{ji} from a selected state i to all other states j, where the weights are the equilibrium populations π_j of the target states j,[70,105,106]

$$\zeta^{orig} \equiv \sum_{n=2}^{N} \tau_n = \sum_{j=1}^{N} \pi_j t_{ji}. \tag{27}$$

Remarkably, the result is independent of the choice of the initial state i.[69]

Since the MFPTs in a Markovian system can be expressed in terms of time-integrated correlation functions,[70] the Kemeny constant can be expressed as the trace of a matrix called "deviation matrix" or "fundamental matrix," which is related to the time-integrated correlation matrix,[107]

$$\zeta^{orig} = \mathrm{Tr}(\hat{\mathbf{C}}\mathbf{D}_\pi^{-1}). \tag{28}$$

The Kemeny constant of the coarse-grained system can be written similarly, and it will depend on the protocol used to coarse-grain the dynamics. For the HS method,

$$\zeta^{HS} = \mathrm{Tr}(\mathbf{S}^T\hat{\mathbf{C}}\mathbf{S}\mathbf{D}_\Pi^{-1}). \tag{29}$$

In Ref. 70, it was shown that the Kemeny constant of the system clustered according to the HS method is bounded by the value of the Kemeny constant in the original system. In particular, the Kemeny constant of the initial and clustered systems are related by a transparent relation, when the HS reduction method is applied,[70]

$$\zeta^{HS} = \zeta^{orig} - \sum_{J=1}^{m} \frac{1}{\Pi_J} \sum_{j \in J, i \in J} \pi_j t_{ji} \pi_i. \tag{30}$$

The second term on the right-hand side (RHS) represents the expectation value of the MFPTs if two states are drawn from within the same cluster with their equilibrium probabilities. This term vanishes when each cluster consists of only one node (i.e., no clustering is performed), and, for a fixed number of clusters, it becomes smaller as the clusters become increasingly metastable, i.e., intra-cluster dynamics is fast as compared to all time scales. The relation above shows that maximizing the Kemeny constant of the clustered system, for a fixed number of clusters, leads to a variationally optimal partitioning, aimed at identifying the clusters with fastest intra-cluster (and the slowest inter-cluster) dynamics.

In conclusion, the Kemeny constant appears to be an objective function that quantifies the metastability of the coarse-grained system, similarly to the modularity, stability, and PCCA+W; however, it has a few advantages when compared to the other measures. One advantage is that it accounts for the information about the system dynamics at all lagtimes, via the integrated correlation function $\hat{\mathbf{C}}$, so that it does not rely on the choice of a particular lagtime τ. In addition, it is variationally optimal and it has a simple interpretations in terms of other kinetic quantities, such as the system timescales and the mean first passage times. Such relations allow for the derivation of explicit formulas for the optimal position of the cluster boundaries

that maximize the Kemeny constant in simple systems, e.g., diffusive processes on 1D potential, as shown in Sec. III A. These features make the Kemeny constant an attractive quantity for the detection of metastable clusters in both, kinetic and unweighted networks.

D. Parallel tempering

Considering the case of clustering a network with N nodes into a coarse-grained one with m nodes, results in m^N possible partitioning. Hence, exact clustering algorithms scale poorly with the size of the system, with most clustering problems known to be NP-hard;[38] therefore, one often uses approximate algorithms to obtain a solution efficiently.

One approach, known as the Louvain method,[92,108] is to iteratively join clusters together until the objective function (modularity) is maximized. Although this approach is fast to execute, it is not readily applicable to many objective functions, such as the Kemeny constant, which always decreases when joining clusters together, following Eq. (30).

Another approach to reducing the complexity of the cluster assignment, consists in defining an ordering for the nodes in the network, i.e., in projecting the nodes onto a one-dimensional space. Subsequently, the task of finding m optimally selected subsets in a network of N nodes, is reduced to finding $m - 1$ separating boundaries. Two spaces that have been used for the projections are the eigenvector associated with the slowest process, and the committor probability with respect to the two most "distant" nodes in the network, often defined as the pair (i, j) with the largest MFPT, t_{ij}. This, however, requires the computation of the MFPTs for all the nodes in the network, which has a large computational cost for large networks.

The problem of extremizing a function dependent on the cluster assignment of the nodes in a network is very common in statistical physics. In particular, assigning a cluster to a node is equivalent to assigning a spin value to a node in the Potts model;[109–111] hence, one can apply methods devised in statistical physics to simulate large physical systems for the task at hand.

Parallel tempering[78,79,112–114] is a simulation method[115] that was developed for the purpose of studying physical systems coupled to a heat bath over extended temperature intervals, in particular, down to small temperatures. It consists in running s parallel simulations of the same system, i.e., different replicas, each at a different temperature[116] and allows replicas to be exchanged between neighboring temperatures. This results in the high temperature replicas exploring the configuration space freely while the low temperature replicas exploring the low energy regions more extensively, hence overall reducing the time to escape from a local minimum of the energy. When concentrating on the sampled configurations of replicas while being at the very low temperatures, the approach was, in particular, used to find the global minimum energy configurations of complex physical systems, for example, for spin glasses, which is NP-hard.[117] Clearly, any objective function of a system to be optimized can be considered as an "energy" which allows one to use parallel tempering as a general-purpose optimization method.[81,82]

Here, we use the algorithm to obtain clusterings; therefore, we use a corresponding notation. In detail, every simulation of a replica α evolves according to a standard Metropolis-Hastings algorithm.

This means, one generates a Markov chain $\mathbf{S}^{(0)} \rightarrow \mathbf{S}^{(1)} \rightarrow \cdots$ of clustering configurations (we omit the α here for brevity), which is not to be confused with the Markov chains of the MSM. For a given clustering $\mathbf{S} = \mathbf{S}^{(l)}$ at "time" l, exhibiting energy $E = E(\mathbf{S}^{(l)})$, a trial configuration $\mathbf{S}'$ is randomly constructed, where the corresponding energy is denoted as E'. Typically $\mathbf{S}'$ is obtained by a random small change to $\mathbf{S}$. The construction probability that $\mathbf{S}'$ is obtained from $\mathbf{S}$ is denoted as $C(\mathbf{S} \rightarrow \mathbf{S}')$. Often this is constant, but not always. Now, the trial configuration will become the next configuration in the Markov chain, i.e., $\mathbf{S}^{(l+1)} = \mathbf{S}'$ with the acceptance probability $p_{acc}^{(\alpha)}$ that depends on the difference between the energies E and E' and on the temperature $T^{(\alpha)}$ of the replica α,

$$p_{acc}^{(\alpha)} = \min\left(1, e^{(E-E')/k_B T^{(\alpha)}} \frac{C(\mathbf{S}' \rightarrow \mathbf{S})}{C(\mathbf{S} \rightarrow \mathbf{S}')}\right), \quad (31)$$

where k_B is the Boltzmann constant. Otherwise, with probability $1 - p_{acc}^{(\alpha)}$, the current configuration will be kept, i.e., $\mathbf{S}^{(l+1)} = \mathbf{S}^{(l)}$. Note that the fraction is one if the construction probability is constant. This choice for the acceptance probability ensures detailed balance between neighboring configurations at the same temperature, hence leads to convergence to the Boltzmann equilibrium measure. A standard time unit in this standard simulation is a sweep. Within a sweep, each degree of freedom is allowed to be changed on average once within a trial configuration.

At regular time intervals, typically after a sweep for each replica is performed, configurations at different temperatures (in different replicas) are interchanged with a probability

$$p_{\alpha\beta} = \min\left(1, e^{(E^{(\alpha)} - E^{(\beta)})\left(\frac{1}{k_B T^{(\alpha)}} - \frac{1}{k_B T^{(\beta)}}\right)}\right), \quad (32)$$

where α and β are the indices of the simulations being interchanged and $E^{(\alpha,\beta)}$ are the energies of the configurations in replicas α and β, respectively. Here again, detailed balance is guaranteed; hence, the simulation will equilibrate with respect to the product measure of two neighboring replicas. In practice, one usually only considers moves between replicas where the temperatures are direct neighbors in the ordered list of temperatures. This applies to all pairs of (neighboring) temperatures, i.e., for the full system.

The temperature set $\{T^\alpha\}$ needs to be specified. A standard rule of thumb is that the empirical acceptance frequency of an exchange between neighboring temperatures should be roughly 0.5. Usually, one performs short test simulations for various sets of temperatures. Thus, if the observed frequency is too small, one would move the respective two temperatures closer to each other, possibly one has to increase the number s of temperatures. For very high acceptance frequencies, it is opposite. Typically, this results in sets where at low temperatures the differences between neighboring temperatures are small, while at high temperatures, the differences are large. In the present work, we follow the same main idea, but use a different protocol as explained in Sec. III B.

III. RESULTS

A. Analytic maximization of the Kemeny constant

In this work we propose to use the Kemeny constant of the coarse-grained dynamics as the objective function to maximize, in order to identify the optimal clustering. Given that the Kemeny constant is related to MFPTs by a simple relation, it turns out that for systems diffusing in a 1D potential, one can calculate analytically the position of the optimal boundaries between clusters.

1. Diffusion in 1D symmetric potential: Three-state clustering

In this section, we consider a 1D system diffusing in a symmetric potential $U(x)$. We first consider clustering the system into three clusters. Since the potential is symmetric, we assume that the position of the boundaries separating the clusters are also symmetric, and positioned at $-a$ and a, where 0 is the center of the 1D space. It is the aim of the clustering to determine the most suitable cluster boundary a.

Considering the definition of the Kemeny constant in Eq. (27), taking the central cluster, denoted "2" and containing the region $(-a, a)$, as the starting state, and using the symmetry of the boundaries, we obtain

$$\zeta^{CG} = 2\Pi_1 t_{12}^{CG}, \quad (33)$$

where Π_1 is the equilibrium occupation probability of cluster 1, which is equal to Π_3 by symmetry, and t_{12}^{CG} is the MFPT from $I = 2$ to $J = 1$ in the coarse-grained system. In Ref. 70, we have derived an expression for t_{JI}^{CG} for discrete-state Markov processes. By replacing summations with integrals in that expression, we can similarly write, for Markov processes in continuous space,

$$t_{JI}^{CG} = \frac{1}{\Pi_I \Pi_J} \int_I dx \int_J dy \quad \pi(y) t_{yx} \pi(x)$$
$$- \frac{1}{\Pi_J^2} \int_J dx \int_J dy \quad \pi(y) t_{yx} \pi(x), \quad (34)$$

where $\pi(x) = e^{-U(x)/kT}/Z$ is the Boltzmann distribution and $Z = \int dx\, e^{-U(x)/kT}$ is the partition function.

Both Π_1 and t_{12}^{CG} are dependent on a, and one can maximize ζ^{CG} with respect to the position of the boundary a, by equating $\partial \zeta^{CG}/\partial a$ to zero, which results in the following relation (see Appendix A):

$$\Pi_1 t_{-aa} - \bar{t}_{a2} + \frac{\Pi_2}{\Pi_1} \bar{t}_{-a1} = 0, \quad (35)$$

where

$$\bar{t}_{xI} = \int_I dy \frac{t_{xy} \pi(y)}{\Pi_I} \quad (36)$$

denotes the MFPT from cluster I to a single position x.

Moreover, we can write (see derivation in Appendix A)

$$\frac{\partial \zeta^{CG}}{\partial a} = 2\pi(a)\left[t_{3a}^{RT} - t_{2a}^{RT}\right], \quad (37)$$

where

$$t_{J\alpha}^{RT} = \bar{t}_{\alpha J} + \hat{t}_{J\alpha} \quad (38)$$

represents the round-trip time[118,119] between the single position of the boundary α and the cluster J, with $\hat{t}_{J\alpha}$ defined in Eq. (A16). This quantity naturally provides a distance metric based on mean

first passage times, as it satisfies the symmetry property $t_{IJ}^{RT} = t_{JI}^{RT}$ and the triangle inequality. Interestingly, the round-trip times for two-state systems have been related to the flux between the two states.[120] Moreover, round-trip times, also called commute times, have been previously considered for clustering[121,122] and applied for image classification.[123]

From Eq. (37), it is clear that the optimal position of the boundary a is such that the round-trip times from the clusters that it separates are equal,

$$\frac{\partial \zeta^{CG}}{\partial a} = 0 \quad \Leftrightarrow \quad t_{3a}^{RT} = t_{2a}^{RT}. \qquad (39)$$

By symmetry, $-a$ makes the round-trip times t_{1-a}^{RT} and t_{2-a}^{RT} equal.

To test the result in Eq. (35) on numerical examples, we consider systems diffusing in 1D double- and triple-well potentials of mean force (PMF), given by

$$U(X) = \cos(\alpha(X + \pi)) + e^{-X^2}, \qquad X \in [-\pi; \pi], \qquad (40)$$

where $\alpha = 2, 3$ is the number of wells (local minima).

To map the continuous motion in 1D to a Markov process on a linear chain, we discretize the configuration space X in N bins, which we label with i. We define the transition rates governing the motion between the bins as $K_{ji} = 0$ for $j \neq i \pm 1$ and

$$K_{(i\pm1)i} = Ae^{-[U(i\pm1)-U(i)]/2k_BT}, \qquad (41)$$

where A is a constant set to 1, and T is the temperature of the system, set to 298 K in our examples.

Then, the MFPT matrix $\mathbf{t}$ is obtained from the transition rates matrix $\mathbf{K}$ using the formula[70]

$$t_{ji} = \frac{1}{\pi_j}\left[\left(\boldsymbol{\pi}\mathbf{1}_n^T - \mathbf{K}\right)_{jj}^{-1} - \left(\boldsymbol{\pi}\mathbf{1}_n^T - \mathbf{K}\right)_{ji}^{-1}\right]. \qquad (42)$$

From these, ζ^{CG} is computed by using the definition of the Kemeny constant $\zeta^{CG} = \sum_{J\neq I}\Pi_J t_{JI}^{CG}$ and Eq. (34) for the coarse-grained MFPTs. In addition, the LHS of Eq. (35), denoted as $\Delta\zeta$, is computed for each value of the boundary position a, using Eq. (42).

Figure 1 shows $\Delta\zeta$ as a function of the boundary position a and the position of the boundaries for which ζ^{CG} is maximized, for $\alpha = 2$ panel (a) and $\alpha = 3$ panel (b), respectively. At these positions, $\Delta\zeta$ vanishes as expected. For comparison, the optimal boundary position resulting from the maximization of the slowest timescale of the clustered system, τ_2^{CG}, is shown in the same plot. We compute τ_2^{CG} for each value of a using the relation[77] $\tau_2^{CG} = \Pi_1 t_{-aa} + \bar{t}_{-a1}$. The boundary position that maximizes τ_2^{CG} is known to satisfy[77]

$$\Pi_1 t_{-aa} - \bar{t}_{-a1} = 0. \qquad (43)$$

The left-hand side (LHS) of Eq. (43), denoted with $\Delta\tau_2$, is seen to vanish at the positions that maximize τ_2^{CG}, as expected.

When the number of potential wells (i.e., metastable states) is lower than the number of clusters, both methods identify a transition cluster, which is broader when the Kemeny constant is maximized. Conversely, when the number of potential wells is greater or equal than the number of clusters, the two methods lead to very
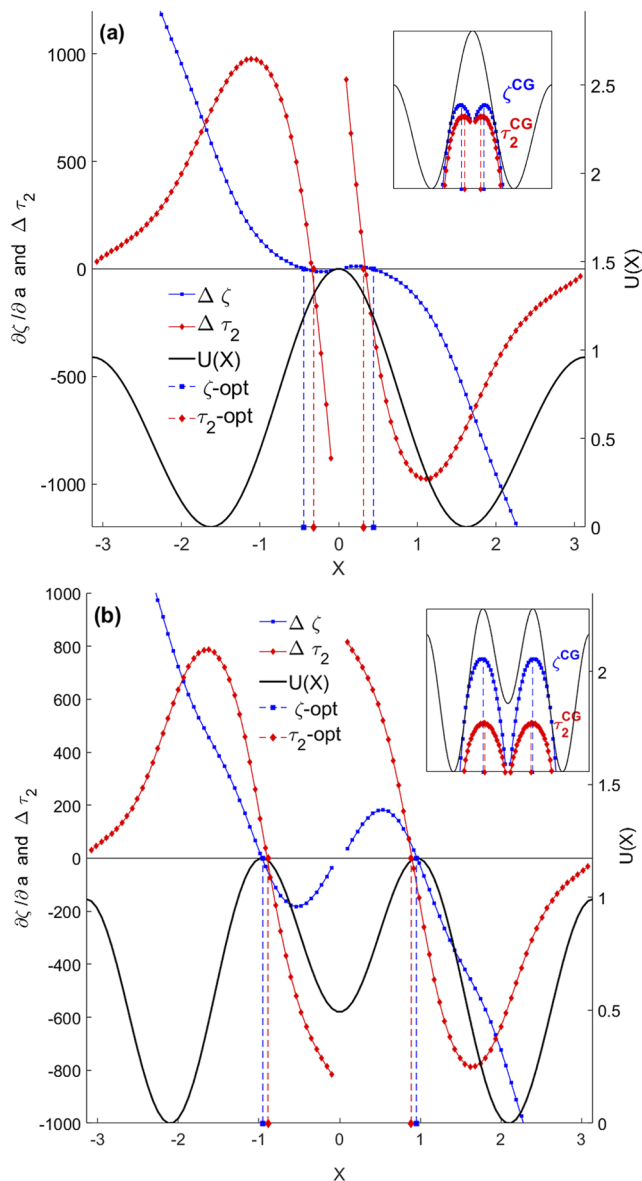
FIG. 1. Three-state clustering of system diffusing on symmetric 1D double well (a) and triple well (b) potentials (black line, y axes on right in kcal/mol units). The dashed lines show the boundary positions corresponding to maxima of the relaxation time τ_2^{CG} (red diamonds, dashed red line) and Kemeny constant ζ^{CG} (blue squares, dashed blue line) in the clustered system. Symbols highlight the end points of these lines. $\Delta\zeta$ (blue curve with squares) defined via the relation in Eq. (35) and $\Delta\tau_2$ (red curve with diamonds) defined via Eq. (43) are plotted against the boundary position. The zeros of $\Delta\zeta$ and of $\Delta\zeta$ yield the corresponding optimum boundary positions, respectively. The inset shows ζ^{CG} (blue squares) and τ_2^{CG} (red diamonds).

similar boundaries positions, which are found around the top of the potential barriers (Fig. 5). The approach presented here assumes a symmetric potential and symmetrically placed optimal boundaries. In Sec. III A 2, we relax these assumptions and develop a more

general method, applicable to an arbitrary number of boundaries and clusters.

2. Diffusion in 1D potential: m-state clustering

Next, we consider coarse-graining a system diffusing in a 1D asymmetric potential into m states. When considering motion in a continuous one-dimensional space, with clusters defined by the positions of the separating barriers b_L, we can compute the derivative of the coarse-grained Kemeny constant with respect to any barrier position b_L,

$$\frac{\partial \zeta^{CG}}{\partial b_L} = \pi(b_L)\left[t_{b_L(L+1)}^{RT} - t_{b_L L}^{RT}\right], \quad (44)$$

where $\pi(x)$ is the probability density function of the position in space, and $t_{J\alpha}^{RT}$ is the round-trip time given in Eq. (38). See the full derivation of the result in Appendix B.

From Eq. (44), it is clear that the position of the barrier b_L maximizes the coarse-grained Kemeny constant when b_L is at equal round-trip distance from cluster L and cluster $L + 1$.

3. Random walk on linear chain: m-state clustering

In the case of a 1D lattice, or linear chair, we can obtain a computationally efficient formula for the finite difference $\Delta\zeta^{CG}$ when moving each boundary position $b_J, J = 1, \ldots, m-1$ from cluster $J + 1$ to J (that is the discrete-space analog of the partial derivatives of the Kemeny constant with respect to the barrier positions). This is found to be (see Appendix C for details)

$$\Delta\zeta^{CG}(b_J, J+1 \to J) = \pi_{b_J} \times \left[\frac{\sum_{S<J}\Pi_S}{\Pi_J^2}\sum_{j\in J}\pi_j t_{jb_J} - \frac{\sum_{S\geq J}\Pi_S}{\Pi_J^2}\sum_{i\in J}\pi_i t_{b_J i}\right.$$
$$\left. + \frac{\sum_{S\leq J+1}\Pi_S}{\Pi_{J+1}^2}\sum_{i\in J+1}\pi_i t_{b_J i} - \frac{\sum_{S>J+1}\Pi_S}{\Pi_{J+1}^2}\sum_{j\in J+1}\pi_j t_{jb_J}\right].$$
$$(45)$$

The optimal boundary positions correspond to the set of boundaries where $\Delta\zeta^{CG}(b_J, J+1 \to J)$ is closest to 0.

To test the results of Eq. (45) on analytical potentials, we first identify the optimal solution by exhaustive search, represented by the vector $\mathbf{b}^* = (b_1^*, b_2^*, \ldots, b_{m-1}^*)$. We then compute the values of $\Delta\zeta^{CG}(b_J, J+1 \to J)$ for all possible positions of b_J between the optimal boundaries b_{J-1}^* and b_{J+1}^*, where the remaining boundaries $\mathbf{b}^{*(J)} = (b_1^* \cdots b_{J-1}^*, b_{J+1}^* \cdots b_{m-1}^*)$ are kept constant. For this purpose, we considered the symmetric three-well potential defined by Eq. (40) with $\alpha = 3$, and an asymmetric four-well potential with barriers of varying height, defined by

$$U(X) = 0.1X(\cos X - 1) \quad \forall X \in [0, 8\pi]. \quad (46)$$

For the symmetric three-well potential [Fig. 2(a)], the boundary positions that maximize ζ are found at the two local maxima of the potential, consistently with results from Fig. 1(a). This is also the case in the asymmetric four-well potential [Fig. 2(b)], where the optimal boundary is identified at the top of the barrier. In both examples, the optimal positions coincide with the locations where the numerical derivative given in Eq. (45) is theoretically predicted to vanish (Fig. 2 colored curves with symbols).
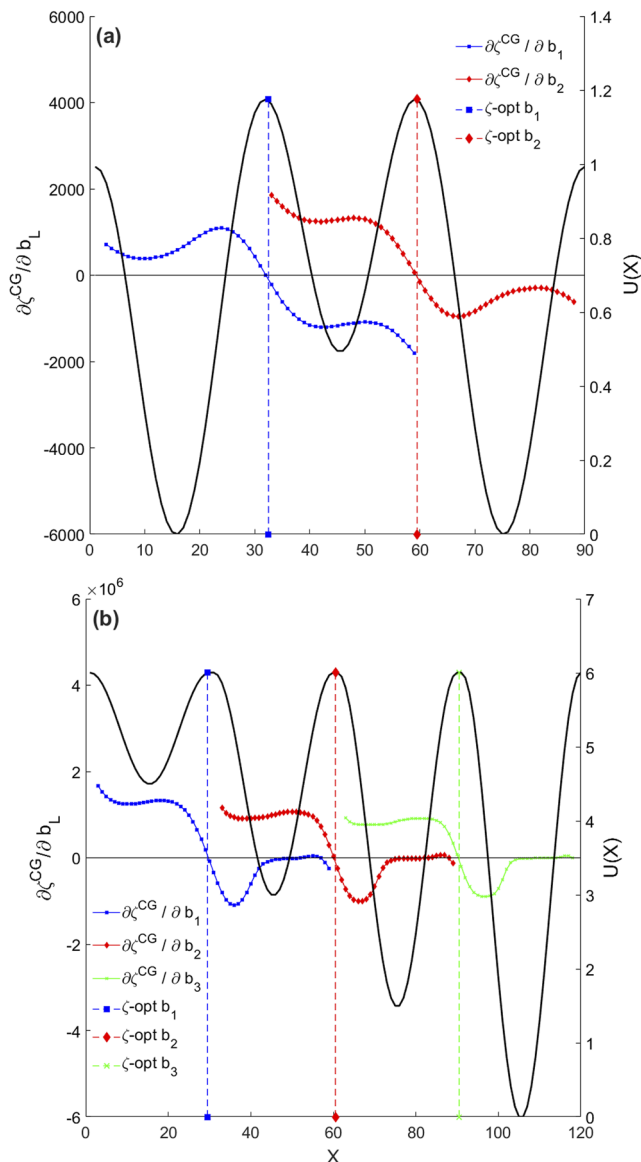
FIG. 2. Three-state (a) and four-state (b) clustering of system diffusing on symmetric (a) and asymmetric (b) 1D potentials (black line, y axes on right in kcal/mol units), respectively. Colored dashed lines show the boundary positions corresponding to the global maximum of ζ^{CG} found by exhaustive search. Dashed lines show $\partial\zeta^{CG}/\partial b_J$ with each color (blue, red, or green) and associated symbol (square, diamond and x) corresponding to a different value of J (1, 2, 3, respectively), were computed using Eq. (45) for all possible positions of $b_J \in (b_{J-1}, b_{J+1})$ when the optimal positions b_l for all $l \neq J$ are kept fixed.

4. Random walks on complex networks: m-state clustering

The result provided in Eq. (44) extends to higher dimensional lattices and complex networks. Several important processes can be modeled as random walks on complex networks. These include diffusive processes in higher dimensions: upon discretizing the

configuration space of such systems in N states, there are in general multiple paths between two states i and j, as described by the links of a networks.

Using the assumption that the population of the selected node is small compared to the cluster populations, $\pi_\alpha \ll \Pi_A$ and $\pi_\alpha \ll \Pi_B$, we find that the change in the coarse-grained Kemeny constant upon moving node α from cluster A to B is given as

$$\Delta\zeta^{CG}(\alpha, A \to B) = \pi_\alpha\left[t_{\alpha A}^{RT} - t_{\alpha B}^{RT}\right], \tag{47}$$

where π_α is the equilibrium probability of node α (see derivation in Appendix D). We note that in contrast to linear chains, in complex networks, there is no natural 1D ordering of nodes; hence, the analytical expressions given by Eq. (45) do not hold as they rely on relations of the type $t_{ba} = t_{bj} + t_{ja} \forall j \in (a, b)$, which are valid only when all transitions between a and b go via the intermediate state j.

Based on our final results from Eq. (47), the optimal assignment of node α will be the closest cluster as measured via the round trip time distance (assuming that the individual population of α is small compared to those of the clusters). Otherwise, we could increase the Kemeny constant by moving α from its current cluster to another one to which its round trip time is smaller. Therefore, in an optimal clustering of a large complex network that maximizes the Kemeny constant, each node belongs to its nearest cluster measured by the round trip time distance.

B. Parallel tempering variational clustering

Here we propose to use a parallel tempering approach for finding the clustering that optimizes an arbitrary objective function.

In the context of network clustering, we replace the energy term E^α in Eqs. (31) and (32) by the objective function one aims to optimize. The temperature T_i^α is a parameter that governs the acceptance probability of a change in clustering assignments, such that the larger T_i^α, the higher the probability of accepting a move that does not improve the objective function.

To generate an initial starting configuration for the parallel tempering algorithm, we generate an easy to obtain clustering. Therefore, we first find a one-dimensional ordering of the nodes of the network. For this purpose, we consider a kinetic process on the network, characterized by the rate matrix $\mathbf{K}$, and we determine the first and the last nodes of the ordering, i and j respectively, as those with the largest MFPT t_{ij}. Although the MFPTs are not symmetric, this definition of first and last nodes, that we will refer to as "extreme nodes" (red nodes in Fig. 3), allows for a meaningful projection as a starting point of our simulation, while remaining applicable to any system. The remaining nodes are then ordered based on their committor probabilities, i.e., the probability to first reach one state before the others. With this 1D ordering of the states, different clusterings are generated randomly, by placing boundaries randomly along the 1D coordinate. The clustering with the highest value of the objective function is selected, among those randomly generated, as the starting configuration for the PTVC algorithm.
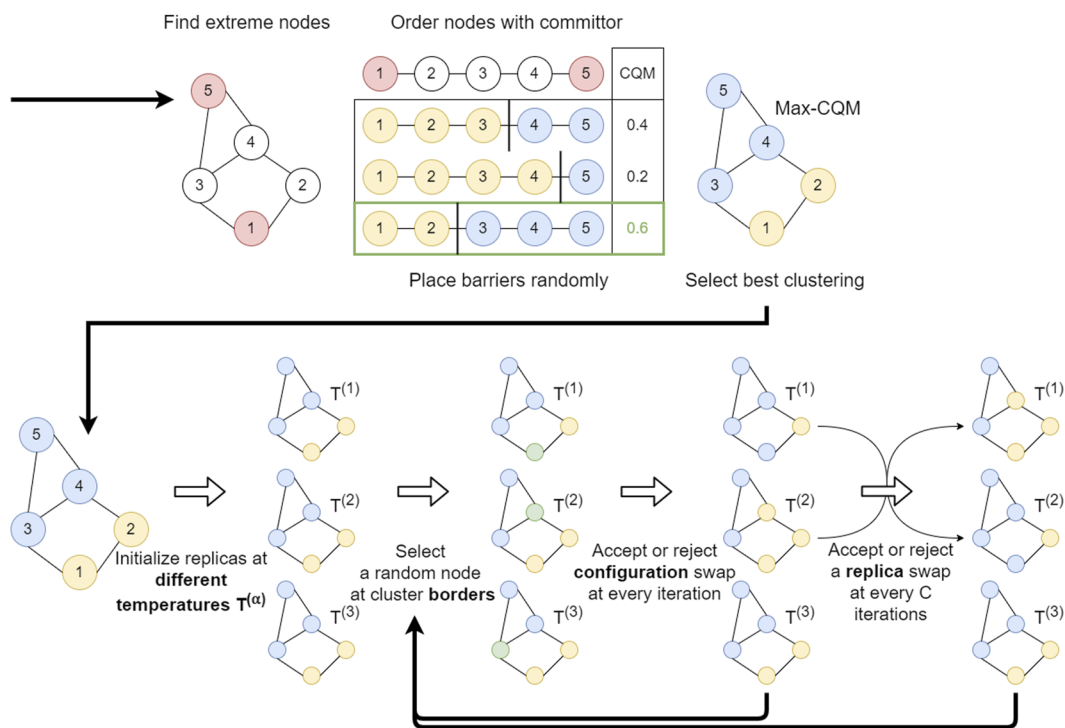
FIG. 3. Illustration of PTVC algorithm. Extreme nodes (having maximal MFPT values) are indicated with red. Clusters are indicated in blue and yellow, and selected bordering nodes are indicated in green.

After initialization of replicas at different temperatures, the PTVC algorithm will no longer be constrained to the 1D projection. Then for each replica, at each time-step of the simulation, we find all the border nodes in the networks and we sample uniformly at random one node i^* from this set, and then among the neighboring clusters of i^*, a cluster J^* is sampled uniformly again. Finally assigning node i^* to cluster J^* is proposed as a trail clustering. We define a node $i \notin J$ as border note of cluster J if there exist a node j such that $A_{ij}S_{jJ} = 1$. We define a cluster I to be neighbor of a cluster J if there are two nodes i, j such that $A_{ij}S_{jJ}S_{iI} = 1$.

The value of the objective function for the clustering after assigning node i^* to cluster J^* is calculated, and the move is accepted or rejected following the acceptance probability in Eq. (31).

The method outlined above (and summarized in Fig. 3) can be applied to arbitrary parameters considered as objective function or energy E. We will use the Kemeny constant and the slowest timescale of the coarse-grained system, respectively, with the purpose of identifying both stable clusters and key transition states.

Other methods in the field of network cluster identification have employed the idea of introducing the concept of temperature to accelerate a variational search through conformations,[84] typically presented as "simulated annealing" methods. These simulated annealing approaches have been shown to find optimized parameter values but are slow. Our method differs from these existing methods in a number of important respects.

Simulated annealing progressively heats and cools the systems to explore configurations. In contrast, parallel tempering runs parallel simulations at multiple temperatures and interchanges configurations at neighboring temperatures.

Our method employs the kinetic timescales of the system as the variational parameter to identify transition states, as opposed to modularity, and we introduce a kinetically motivated initial ordering of the states to enhance the quality of the starting clustering.

C. PTVC algorithm implementation details

Following Eq. (31), the Metropolis–Hastings acceptance probability shall be $p^{\alpha}_{\text{acc}} = \min\left(1, e^{(E-E')/kT^{(\alpha)}} \frac{C(\mathbf{S}' \to \mathbf{S})}{C(\mathbf{S} \to \mathbf{S}')}\right)$ to ensure detailed balance, i.e., equilibrium sampling over all temperatures $T^{(\alpha)}$. Usually, the number of border nodes and neighbor clusters will differ before and after the proposed move, hence the term $\frac{C(\mathbf{S}' \to \mathbf{S})}{C(\mathbf{S} \to \mathbf{S}')}$ is not constant. Nevertheless, we set this term to 1 in our applications as the preservation of detailed balance is not required for the optimization task, hence lowering the computational cost.

One practical consideration that requires discussion is the choice of temperature to be used, which has no physical motivation here and exists only to control the probability of accepting proposed switches for different simulations. To determine the temperature to use, we propose to initialize the s simulations at equally spaced increasing temperatures values $T = T^{(i)}$, where $i \in \{1, \ldots, s\}$, with $T^{(1)} < T^{(2)} < \cdots < T^{(s)}$. We note that it is also possible to optimize the temperature range and spacing using various objectives,[124,125] e.g., to achieve faster round trip times or possibly more relevant relaxation times in the system.[126,127] In our applications presented here, we have used $s = 50$ simulations, and initial temperatures bounded by $T^{(1)} = 0.001$ and $T^{(s)} = 1$.

This effectively enforces that, at first, mostly the proposed moves that optimize the parameter are accepted. From here, we can then increase the temperatures until the average acceptance probability reaches a desired value, which we set to 50%. We achieve this by the computing the proportion of accepted moves since the latest temperature update (p^a) every ten sweeps and then update the temperatures using the formula $T^{new} = T \times \frac{\log(p^a)}{\log(0.5)}$ This ensures that temperatures are high enough for the system to explore the whole configuration space, but not high enough to cause the algorithm to accept every proposed move.

The value of the objective functions was monitored over simulations of 20 000 sweeps, with $s = 50$ different temperatures, and the clustering with the highest objective function was chosen.

D. Clustering results

We use the PTVC algorithm in five test systems to maximize three different parameters: (i) the modularity [Eq. (8)], (ii) the Kemeny constant [Eq. (30)], and (iii) the slowest timescale τ_2 (Sec. II C 4). We compare the resulting clusterings with that of PCCA+. Our test systems consist of a diffusion process in a 1D potential, a synthetic network generated from the stochastic block model (SBM)[128] and three real-world networks.

1. One-dimensional energy profile

Before applying the method to complex networks, we test the PTVC method on a simple 1D model, where the results can be easily interpreted and compared to an exhaustive search for the global maximum.

We consider a potential with four wells of varying depth, defined by Eq. (S1). When clustering into three states, we obtain identical clusters for PCCA+, the Kemeny constant and τ_2 clustering (Fig. 4). These correspond to the three most stable local minima
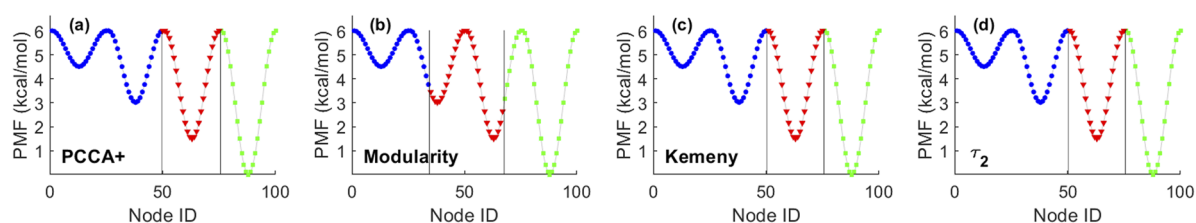
FIG. 4. Clustering a 1D multi-well potential into three clusters (blue circles, red triangles, and green squares symbols with black vertical lines showing the boundaries) using PTVC with the three different target functions: modularity (b), the Kemeny constant (c), and τ_2 (d), respectively. PCCA+ clustering (a) also shown for comparison.
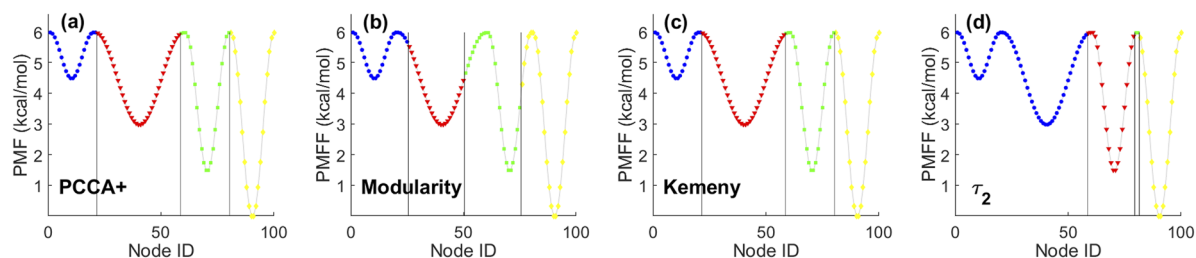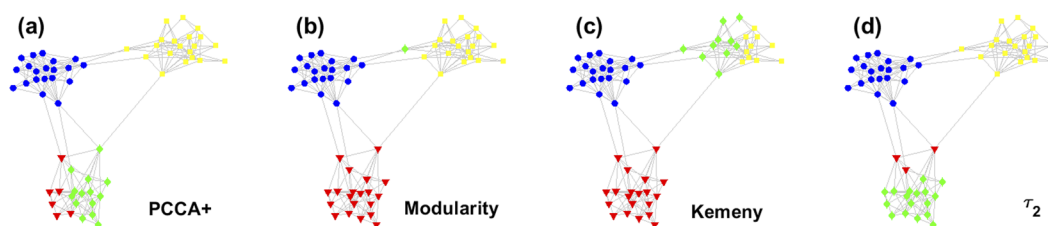
FIG. 5. Clustering a 1D multi-well potential into four clusters (blue circles, red triangles, and green squares symbols with black vertical lines showing the boundaries) using PTVC with the three different target functions: modularity (b), the Kemeny constant (c), and τ_2 (d), respectively. PCCA+ clustering (a) also shown for comparison.

on the potential energy profile separated by the top of the barriers. However, modularity clusters the coordinate space into three equal clusters, independently on the underlying free energy profile. This is expected as all networks corresponding to a 1D profile have the same tridiagonal connectivity.

The optimal four-state clustering on the same potential (Fig. 5 and Table I) also results in the same clusters for PCCA+ and the Kemeny constant, identifying all four local minima. Analogously as for three-state clustering, modularity identifies an equal spacing of the reaction coordinate into four parts, which does not take into account the free energy profile. Interestingly, however, τ_2 clustering results in a fourth state that is a transition state (TS), suggesting that the highest energy local minimum is less important for the slowest timescale of the process, and a cluster with the TS node is more optimal to maximize τ_2. Therefore, the number of metastable clusters can be determined by observing the first time the maximization

of ζ^{CG} and τ_2^{CG} yield different results. Analogously, the eigenvalue spectrum also suggests three metastable states for this system [Fig. S1(a)]. In general, once we move beyond the number of metastable states, the Kemeny constant identifies smaller metastable states, and τ_2 finds a cluster with very small population around the dominant barrier.

2. Stochastic block model

To analyze more complex networks beyond the one-dimensional connectivity, we generated three-state random networks using the stochastic block model (SBM) (Fig. 6). The adjacency matrix of a random SBM network is constructed according to

$$P(A_{ij} = 1) = \frac{cW(X_i, X_j)}{NP(X_i)P(X_j)}, \qquad (48)$$

where $X_i \in 1, \ldots, m$ indicates the cluster of node i, $P(X)$ is the probability of a node being in cluster X, $W(X, X')$ is the probability of a link existing between cluster X and X', N is the total number of nodes in the network, and $c = 4$ is the average connectivity of the network. In our applications, we have used the uniform distribution for $P(X) = 1/m$, with $m = 3$ clusters, $N = 99$ nodes in total, and intra-cluster and inter-cluster edge existence probability of $W(X, X) = 0.7$ for all X and $W(X, X') = 0.005$ for all $X \neq X'$, respectively.

For SBM networks, there is no notion of dynamics, hence we consider a random walk on the network, with the Laplacian introduced in Eq. (6). Therefore, the dynamics of an unbiased walker is characterized by frequent intra-cluster transitions and rare inter-cluster transitions so that the network communities correspond to the clusters as metastable states.

TABLE I. One-dimensional potential energy surface reduced to four clusters as shown in Fig. 5. Values of each objective function: the Kemeny constant, τ_2 and modularity (columns) for each clustering method PCCA+, objective functions Kemeny constant, τ_2 and modularity (rows) are shown. The largest value for each objective function is highlighted in bold.

Clustering method	Kemeny	τ_2	Modularity
PCCA+	**53 185**	41 780.6	0.417 16
Kemeny	**53 185**	41 780.6	0.417 16
τ_2	52 308	**41 781.3**	0.380 43
Modularity	52 865	41 660.2	**0.422 34**

FIG. 6. Clustering of a three-state stochastic block model network into four clusters using PTVC with the three different target functions: modularity (b), the Kemeny constant (c), and τ_2 (d). PCCA+ clustering (a) also shown for comparison.

TABLE II. Values of the Kemeny constant, τ_2 and modularity for a three-state stochastic block model network reduced to four clusters using the PCCA+ method, and the optimization for the Kemeny constant, τ_2, and modularity. Bolded numbers indicate the largest value of the respective objective function.

Clustering method	Kemeny	τ_2	Modularity
PCCA+	65.342	40.774	0.364 5
Kemeny	**65.574**	40.807	0.359 08
τ_2	65.373	**40.942**	0.378 32
Modularity	65.091	40.796	**0.396 52**

TABLE III. Values of the Kemeny constant, τ_2 and modularity for the Santa Fe collaboration network reduced to five clusters using the PCCA+ method, and the optimization for the Kemeny constant, τ_2 and modularity. Bolded numbers indicate the largest value of the respective objective function.

Clustering method	Kemeny	τ_2	Modularity
PCCA+	335.58	238.51	0.417 49
Kemeny	**336.2**	238.8	0.415 04
τ_2	333.11	**239.9**	0.416 52
Modularity	334.53	236.52	**0.419 99**

We used the PTVC algorithm to identify three clusters within a three-state SBM network for all three variational parameters. As expected, we obtain the correct clustering using all three methods as well as using PCCA+ (Fig. S2 and Table S1).

The results for four clusters (Fig. 6 and Table II) show a consistent picture with the 1D example. PCCA+ and Kemeny subdivides one of the clusters in two approximately equal-sized clusters, whereas τ_2-optimal clustering identifies a transition cluster separating one small metastable cluster from the others. Modularity identifies the three expected clusters and assigns a single node to an additional fourth cluster.

Moreover, we find that the identified TS contains both the nodes with highest closeness and betweenness centrality (Fig. S3), which indeed reflects the nature of transition states: they are positioned *between* metastable states, most of the paths go through them, and they are *close* to all the other nodes.

3. Santa Fe collaboration network

As a first example of a real world network that is small enough for clear visualization, we consider the Santa Fe collaboration network (Fig. 7).[129] In this network, links are drawn between 118 researchers at the Santa Fe institute who appeared as co-authors on at least one publication. The nodes then form clusters corresponding to three main research groups, linked by cross-disciplinary researchers.

We apply random-walk dynamics on this network and cluster it into three states (Fig. S4 and Table S2), which corresponds to the number of metastable clusters predicted by the spectral gap (Fig. S1), we find that the resulting clusterings are identical for all four methods, as in the previously considered networks.

When clustering into four states (Fig. S4 and Table S3), we find that the τ_2-optimal clustering still finds a transition cluster separating two of the metastable regions as expected. However, the Kemeny-optimal clustering no longer behaves similarly to PCCA+ clustering and is instead identical to the τ_2-optimal one.

The fact that the Kemeny and τ_2-optimal clusterings are identical can be seen as a consequence of the fact that the additional cluster is also metastable but too small to be detected via spectral gap; hence, it is insightful to look for one more cluster. We find indeed that when optimizing for five clusters (Fig. 7 and Table III), the Kemeny-optimal and τ_2-optimal clusterings are no longer identical, and the latter successfully identifies the two regions separating the three initial metastable states into distinct clusters. Again, we find that the two additional clusters contain nodes with high closeness and betweenness centrality (Fig. S5).

In conclusion, our method retrieves all the communities when the number of clusters used coincide with the number of communities detected via spectral gap analysis, and it allows us to identify transition clusters containing the nodes with the highest centrality when used with a larger number of clusters.

4. Political books co-purchasing network

Our next real-world network is one of political books: each node represents a book, and an edge is present between two books if they appear as being often co-purchased on Amazon. Thus, for our clustering approach, we also use the random-walk dynamics. The data have been compiled by Valdis Krebs.[130] Based on a reading of the descriptions and reviews of the books on Amazon, the nodes have been labeled manually by Mark Newman into one of three categories: "liberal" (blue), "neutral" (red), and "conservative" (green)[131] (Fig. S6), providing a comparison to our clusterings.
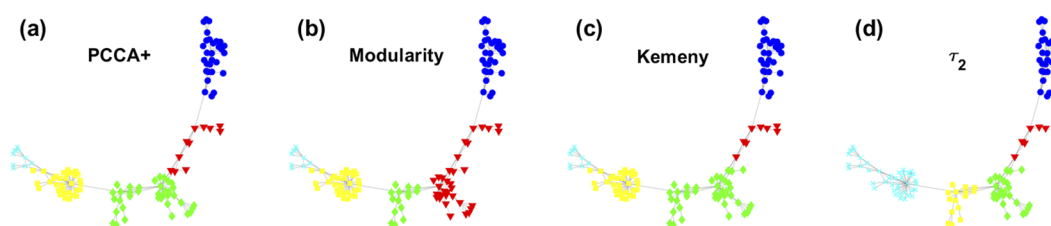
FIG. 7. Clustering of the Sante Fe Institute collaboration network into five clusters using PTVC with the three different target functions: modularity (b), the Kemeny constant (c), and τ_2 (d). PCCA+ clustering (a) also shown for comparison.
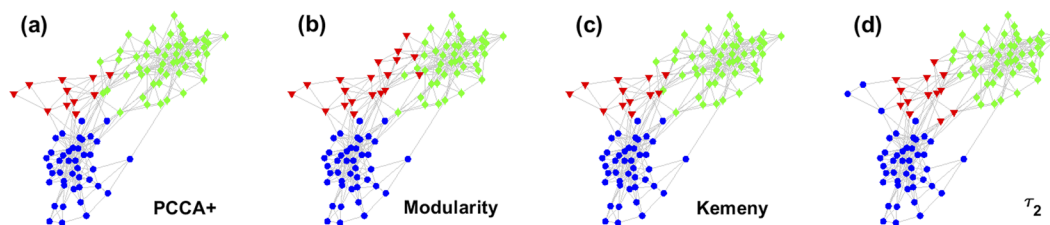
FIG. 8. Clustering of the political books co-purchase network into three clusters using PTVC with the three different target functions: modularity (b), the Kemeny constant (c), and τ_2 (d). PCCA+ clustering (a) also shown for comparison. The "conservative" and "liberal" books clusters (blue and green) are similarly detected by all parameters, with differences mainly on the "transition" cluster (red).

TABLE IV. Values of the Kemeny constant, τ_2 and modularity for the political book network reduced to three clusters using the PCCA+ method, and the optimization for the Kemeny constant, τ_2 and modularity. Bolded numbers indicate the largest value of the respective objective function.

| Clustering method	Kemeny	τ_2	Modularity
PCCA+	29.944	25.041	0.358 79
Kemeny	**29.954**	25.237	0.358 43
τ_2	29.023	**25.653**	0.347 39
Modularity	29.524	25.062	**0.359 54**

The spectral gap analysis detects two metastable states (Fig. S1), corresponding to the categories "liberal" and "conservative" as we could expect, while the "neutral" category will be considered as the transition state between the two. We observe that partitioning into two clusters yields identical clusterings for all dynamical clustering methods and a very similar clustering for modularity (Fig. S7 and Table S4). When optimizing for three clusters yields results consistent with our previous observations (Fig. 8 and Table IV). On one hand the Kemeny constant and PCCA+ converge to a similar clustering. On the other hand the τ_2-optimal clustering is seen to separate the two stable states (Fig. 8), while the small transition cluster contains nodes with high closeness and betweenness centrality (Fig. S8). Modularity does not separate the two stable states fully and identifies a third cluster somewhat different from Kemeny and PCCA+.

The manual labeling (Fig. S6) is not retrieved by any of the kinetic clusterings we have tried. This can be explained by multiple factors: the edges here are binary, i.e., they do not measure *how strongly* connected (co-purchased) two books are; hence, all the information about the strength of the connection is lost. Given that our method is based on the dynamics of the network, which is highly dependent on the weights of the edges (transition rates), we can expect it to lead to a different clustering in the presence of weights, which may be closer to the manual clustering with a weighted network. Additionally, the manual labeling have been made by a single person, and therefore it is prone to the subjectivity of this person's judgment. The discrepancy between the τ_2-optimal clustering and the labeling can be used to suggest reconsideration of the labeling or to gather a more robust dataset for the task at hand. Furthermore, it also suggests that there is no completely neutral opinion, but some bias exists toward liberal or conservative orientation.

5. Streets network

Finally, we consider a subset of the open street map road networks of Luxembourg (Fig. 9). The graph is an undirected and unweighted version of the largest strongly connected component of
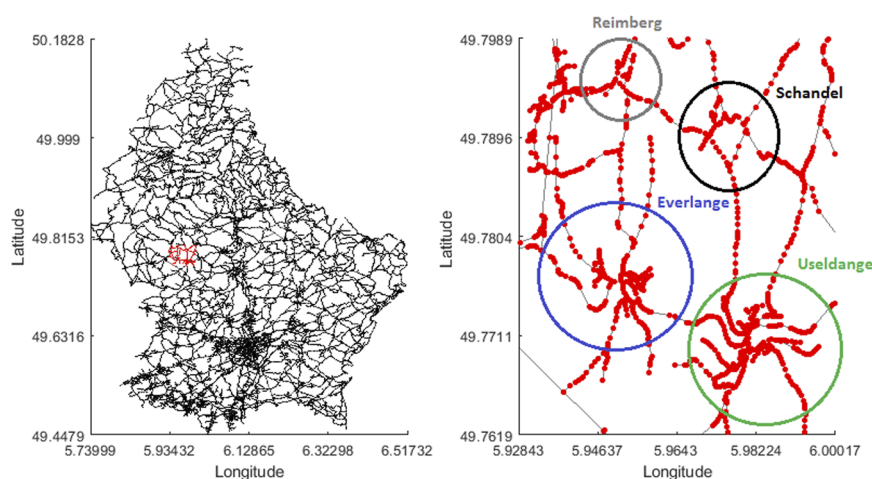
FIG. 9. Street network of Luxembourg from DIMACS10,[136,132] and the subset we consider. Latitude and longitude in decimal degrees; the colored circles highlight the main cities in this geographic area.
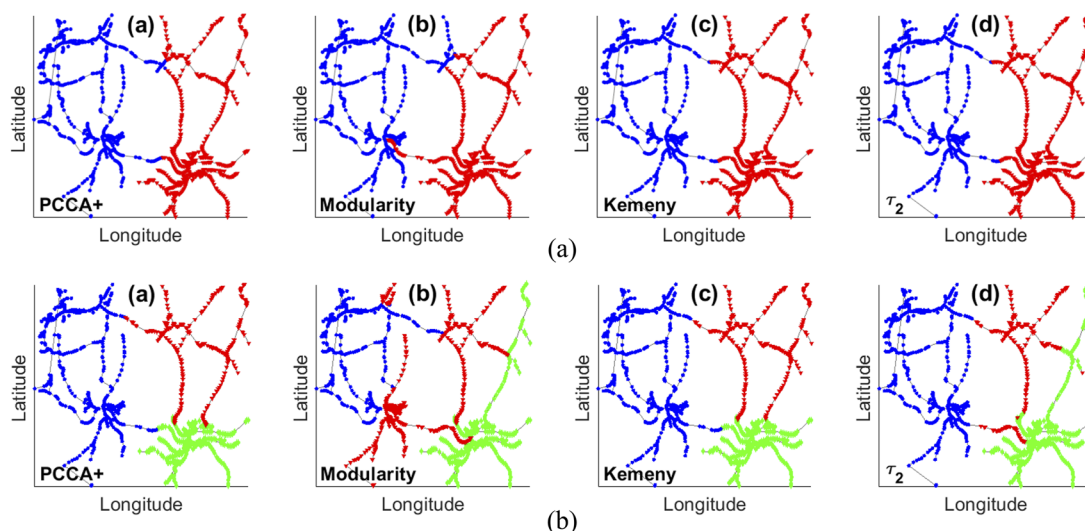
FIG. 10. Clustering of the subset of Luxembourg streets network (Fig. 9, right) into two (top) and three (bottom) states. PCCA+ clustering (a), modularity (b), the Kemeny constant (c), and τ_2 (d) optimized clusters are shown for comparison.

the corresponding open street map road networks.[132] Due to the large size of the network (114 599 nodes), we have taken a subset of the initial graph (Fig. 9, right), the subset contains two medium-sized cities (Everlange and Useldange), as well as a few smaller cities (Reimberg and Shandel).

In the absence of data about the dynamics in this network, we applied the Laplacian method described in Sec. II A 2 to obtain the random walk dynamics. We find that the spectral gap analysis determines that there are two metastable states.

Similarly to what we observe in the previous networks, when optimizing for two clusters, the partitions from all four methods provide similar results (Fig. 10, top and Table V). Each of the two clusters contains one of the two largest cities of the network, Everlange and Useldange (Fig. 9, right).

When clustering into three states (Fig. 10, bottom and Table VI), we find that PCCA+ provides a sensible clustering, identifying the third largest city, Schandel, into the third cluster, similarly to the Kemeny-optimal clustering. The modularity-optimal clustering splits the left-hand cluster, we can see that the additional cluster corresponds to a part of one of the largest cities (Everlange). The τ_2-optimal clustering maintains the two main cities unseparated, and instead assigns the third largest city as well as the main road connecting the two main cities into the third, transition cluster, thus completely separating the two main cities.

TABLE V. Values of the Kemeny constant, τ_2 and modularity for the Luxembourg roads network reduced to two clusters using the PCCA+ method, and the optimization for the Kemeny constant, τ_2 and modularity. Bolded numbers indicate the largest value of the respective objective function.

Clustering method	Kemeny	τ_2	Modularity
PCCA+	14 327	14 327	0.373 65
Kemeny	**14 593**	**14 593**	0.374 04
τ_2	**14 593**	**14 593**	0.374 04
Modularity	13 280	13 280	**0.373 67**

TABLE VI. Values of the Kemeny constant, τ_2 and modularity for the Luxembourg roads network reduced to three clusters using the PCCA+ method, and the optimization for the Kemeny constant, τ_2 and modularity. Bolded numbers indicate the largest value of the respective objective function.

Clustering method	Kemeny	τ_2	Modularity
PCCA+	19 973	15 007	0.405 7
Kemeny	**20 007**	15 035	0.404 99
τ_2	18 167	**15 374**	0.399 9
Modularity	15 260	13 144	**0.413 05**

IV. CONCLUSIONS

Network clustering is a crucial component of analysis of large datasets in most fields. Here, we present a theoretical framework for clustering based on the dynamical properties of networks. We introduce variational clustering protocols using the Kemeny constant and the slowest relaxation time. Building upon our earlier computational and theoretical work on the Kemeny constant,[70,76,77] we derived a novel theoretical expression for the gradient of the Kemeny constant with respect to clustering, and provided an analytical solution for the optimal clustering of a 1D potential maximizing the Kemeny constant, for an arbitrary number of clusters. We showed that boundaries in optimal clustering correspond to equal round-trip times between clusters they separate. We further extended this

result to complex networks, in general, and showed that in the optimal clustering maximizing the Kemeny constant each node belongs to the cluster that is the closest as measured by the round-trip time distance measure.

To enable more efficient clustering in complex systems, we described in the present work an algorithmic protocol, PTVC, allowing for the identification of clusters in large and complex networks by using parallel tempering to maximize any variational parameter or objective function. Using this algorithm, we compared different objective functions (modularity, Kemeny constant, and slowest timescale τ_2) as well as PCCA+ on the examples of two model networks and three real data-derived networks.

The optimization using the Kemeny constant is shown to successfully identify the key metastable states, in both weighted and unweighted networks, illustrating the proof of the variational result.[70] Moreover, maximizing τ_2 has shown to consistently identify less stable states, which can be considered as transition clusters, in addition to the metastable states. These transition clusters appear to contain nodes with high closeness and betweenness centrality. While the vast majority of clustering methods focus on detecting communities (metastable states), we present results suggesting that the use of specific timescales τ_i can be used to also efficiently identifying key transition states in complex systems.

This new approach of maximizing specific timescales (τ_i) or their sum (Kemeny) opens up new and exciting potential research avenues and applications. For one, this method can effectively identify key transition states from a Markov model. In the context of molecular simulations, this can be incorporated to analyze and enhance the sampling of the system's configuration space.[133]

Furthermore, spectral properties are related via the Kemeny constant to mean first passage times, which can be more readily available from numerical data, and can be evaluated numerically effectively.[134–136] The novel analytical expression for the gradient of the Kemeny constant, $\partial \zeta^{CG}/\partial b_J$, and its discrete form using round trip times, derived in this work, also opens the possibilities to employ alternative implementations of our clustering method, based on, e.g., gradient descent techniques, thus avoiding an exhaustive search over the space of possible boundary positions. This will be explored in future work.

SUPPLEMENTARY MATERIAL

The supplementary material contains the definition of the irregular 1D potential; numerical values for the Kemeny constant, τ_2, and modularity for the various networks clustered into different number of clusters; spectral gaps; additional clustering results including centrality measures.

ACKNOWLEDGMENTS

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Vladimir Koskin: Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Adam Kells**: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Writing – original draft (equal); Writing – review & editing (equal). **Joe Clayton**: Data curation (equal); Formal analysis (equal); Writing – original draft (equal). **Alexander K. Hartmann**: Conceptualization (equal); Formal analysis (equal); Writing – review & editing (equal). **Alessia Annibale**: Conceptualization (equal); Formal analysis (equal); Methodology (equal); Project administration (equal); Supervision (equal); Validation (equal); Writing – original draft (equal); Writing – review & editing (equal). **Edina Rosta**: Conceptualization (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

APPENDIX A: DIFFUSION ON 1D SYMMETRIC POTENTIAL: THREE-STATE CLUSTERING

In this section, we consider a system diffusing in a 1D potential symmetric about 0. We consider clustering the system in three clusters, which we shall denote with 1, 2, 3. Taking 0 as the center of the 1D space, we denote with a the position of the boundary between cluster 2 and 3 and, due to the symmetry of the potential, $-a$ will denote the boundary between 1 and 2.

The clustered mean first passage times are given in Eq. (34) and can be written compactly as

$$t_{JI}^{CG} = \frac{1}{\Pi_J} \int_J dx\, \pi(x) [\bar{t}_{xI} - \bar{t}_{xJ}] \tag{A1}$$

in terms of the MFPTs to the single position x from the clusters I and J, respectively, as defined in Eq. (36).

Inserting this into the expression for the Kemeny constant provided for a symmetric system coarse-grained into three clusters, in Eq. (33), we obtain

$$\zeta^{CG}/2 = \int_{-\infty}^{-a} dx\, \pi(x) (\bar{t}_{x2} - \bar{t}_{x1}). \tag{A2}$$

Differentiating with respect to a using the Leibniz integral rule yields

$$\frac{\partial \zeta^{CG}/2}{\partial a} = -\pi(-a)(\bar{t}_{-a2} - \bar{t}_{-a1}) + \int_{-\infty}^{-a} dx\, \pi(x) \frac{\partial}{\partial a}(\bar{t}_{x2} - \bar{t}_{x1}). \tag{A3}$$

From Eq. (36), we have

$$\bar{t}_{x2} = \int_{-a}^{a} dz\, p_2(z) t_{xz} \tag{A4}$$

with

$$p_2(x) = \frac{\pi(x)}{\Pi_2}, \qquad \Pi_2 = \int_{-a}^{a} dx\, \pi(x), \tag{A5}$$

and hence

$$\frac{\partial}{\partial a} \bar{t}_{x2} = p_2(a) t_{xa} + p_2(-a) t_{x-a} + \int_{-a}^{a} dz\, \frac{\partial}{\partial a} p_2(z) t_{xz}. \tag{A6}$$

Substituting

$$\begin{aligned}
\frac{\partial}{\partial a} p_2(z) &= \frac{\partial}{\partial a} \frac{\pi(z)}{\Pi_2(a)} = -\frac{\pi(z)}{\Pi_2^2(a)} \frac{\partial}{\partial a} \Pi_2(a) \\
&= -\frac{p_2(z)}{\Pi_2(a)} \frac{\partial}{\partial a} \Pi_2(a) \\
&= -[p_2(a) + p_2(-a)] p_2(z),
\end{aligned} \tag{A7}$$

we get

$$\frac{\partial}{\partial a} \bar{t}_{x2} = p_2(a) t_{xa} + p_2(-a) t_{x-a} - [p_2(a) + p_2(-a)] \bar{t}_{x2}. \tag{A8}$$

Similarly, for $\bar{t}_{x1} = \int_{-\infty}^{-a} dz\, p_1(z) t_{xz}$,

$$\frac{\partial}{\partial a} \bar{t}_{x1} = p_1(-a)(\bar{t}_{x1} - t_{x-a}). \tag{A9}$$

Collecting all the terms, and using $p_2(a)\pi(x) = p_2(x)\pi(a)$ and similarly for $-a$,

$$\begin{aligned}
\frac{\partial}{\partial a} \zeta^{CG}/2 = &-\pi(-a)(\bar{t}_{-a2} - \bar{t}_{-a1}) + \pi(a) \int_{-\infty}^{-a} dx\, p_2(x) t_{xa} \\
&+ \pi(-a) \int_{-\infty}^{-a} dx\, p_2(x) t_{x-a} \\
&- [\pi(a) + \pi(-a)] \int_{-\infty}^{-a} dx\, p_2(x) \bar{t}_{x2} \\
&- \pi(-a) \int_{-\infty}^{-a} dx\, p_1(x)(\bar{t}_{x1} - t_{x-a}).
\end{aligned} \tag{A10}$$

Using $\pi(a) = \pi(-a)$, due to the symmetry of the potential, and that $\forall\, x \in 1\ \bar{t}_{x2} = t_{x-a} + \bar{t}_{-a2}$ and $t_{xa} = t_{x-a} + t_{-aa}$,

$$\begin{aligned}
\frac{1}{\pi(a)} \frac{\partial}{\partial a} \zeta^{CG}/2 = &-(\bar{t}_{-a2} - \bar{t}_{-a1}) + \int_{-\infty}^{-a} dx\, p_2(x)(t_{x-a} + t_{-aa}) \\
&+ \int_{-\infty}^{-a} dx\, p_2(x) t_{x-a} - 2\int_{-\infty}^{-a} dx\, p_2(x)(t_{x-a} + \bar{t}_{-a2}) \\
&- \int_{-\infty}^{-a} dx\, p_1(x)(\bar{t}_{x1} - t_{x-a}).
\end{aligned}$$

Simplifying, and using $\Pi_1 = \Pi_3$ due to the symmetry of the potential,

$$\begin{aligned}
\frac{1}{\pi(a)} \frac{\partial}{\partial a} \zeta^{CG}/2 = &-\bar{t}_{-a2} + \bar{t}_{-a1} + \frac{\Pi_1}{\Pi_2} t_{-aa} - 2\frac{\Pi_1}{\Pi_2} \bar{t}_{-a2} \\
&- \int_{-\infty}^{-a} dx\, p_1(x)(\bar{t}_{x1} - t_{x-a}) \\
= &-\left(1 + 2\frac{\Pi_1}{\Pi_2}\right) \bar{t}_{-a2} + \bar{t}_{-a1} + \frac{\Pi_1}{\Pi_2} t_{-aa} \\
&- \int_{-\infty}^{-a} dx\, p_1(x)(\bar{t}_{x1} - t_{x-a}) \\
= &-\frac{1}{\Pi_2} \bar{t}_{-a2} + \bar{t}_{-a1} + \frac{\Pi_1}{\Pi_2} t_{-aa} \\
&- \int_{-\infty}^{-a} dx\, p_1(x)(\bar{t}_{x1} - t_{x-a}).
\end{aligned}$$

Equating $\partial \zeta^{CG}/\partial a$ to zero, we obtain

$$-\frac{1}{\Pi_2} \bar{t}_{-a2} + \bar{t}_{-a1} + \frac{\Pi_1}{\Pi_2} t_{-aa} = \int_{-\infty}^{-a} dx\, p_1(x)(\bar{t}_{x1} - t_{x-a}).$$

Because of symmetry, this is equivalent to

$$-\frac{1}{\Pi_2} \bar{t}_{a2} + \bar{t}_{a3} + \frac{\Pi_3}{\Pi_2} t_{-aa} = \int_{-\infty}^{-a} dx\, p_1(x)[\bar{t}_{x1} - t_{x-a}].$$

One can write the right-hand side (RHS) in terms of the Kemeny constant

$$\begin{aligned}
\zeta^{CG}/2 &= \int_{-\infty}^{-a} dx\, \pi(x)(\bar{t}_{x2} - \bar{t}_{x1}) \\
&= \Pi_1 \int_{-\infty}^{-a} dx\, p_1(x)[\bar{t}_{x2} - \bar{t}_{x1}] \\
&= \Pi_1 \int_{-\infty}^{-a} dx\, p_1(x)[(t_{x-a} + \bar{t}_{-a2}) - \bar{t}_{x1}] \\
&= \Pi_1 \int_{-\infty}^{-a} dx\, p_1(x)[t_{x-a} - \bar{t}_{x1}] + \Pi_1 \bar{t}_{-a2} \tag{A11}
\end{aligned}$$

to get

$$-\frac{1}{\Pi_2} \bar{t}_{a2} + \bar{t}_{a3} + \frac{\Pi_3}{\Pi_2} t_{-aa} = \bar{t}_{-a2} - \frac{\zeta^{CG}}{2\Pi_1} = \bar{t}_{-a2} - t_{12}^{CG}, \tag{A12}$$

where, in the last step, we used Eq. (33), or, using equivalence of states 1 and 3,

$$\Pi_1 t_{-aa} - \bar{t}_{a2} + \Pi_2(\bar{t}_{-a1} - \bar{t}_{-a2} + t_{12}^{CG}) = 0. \tag{A13}$$

Another identity can be obtained by using the relation of the quantities above to the Kemeny constant of the original process $\zeta^{orig} = \int_{-\infty}^{\infty} dx\, \pi(x) t_{xy}$. Writing

$$\begin{aligned}
\frac{1}{\Pi_1} \int_{-\infty}^{-a} dx\, \pi(x) \bar{t}_{x1} &= \frac{1}{\Pi_1^2} \int_{-\infty}^{-a} dx\, \pi(x) \int_{-\infty}^{-a} dy\, \pi(y) t_{xy} \\
&= \frac{1}{\Pi_1^2} \int_{-\infty}^{-a} dy\, \pi(y)\left(\zeta - \int_{-a}^{\infty} dx\, \pi(x) t_{xy}\right) \\
&= \frac{1}{\Pi_1}\left(\zeta - \int_{-a}^{\infty} dx\, \pi(x) \bar{t}_{x1}\right) \tag{A14}
\end{aligned}$$

and using Eq. (A1) and $\bar{t}_{x2} = t_{x-a} + \bar{t}_{-a2}\ \forall\, x \in 1$, we obtain

$$
\begin{aligned}
t_{12}^{CG} &= \int_{-\infty}^{-a} dx \, \frac{\pi(x)}{\Pi_1} \left(t_{x-a} + \bar{t}_{-a2} - \bar{t}_{x1} \right) \\
&= \int_{-\infty}^{-a} dx \, \frac{\pi(x)}{\Pi_1} t_{x-a} + \bar{t}_{-a2} - \int_{-\infty}^{-a} dx \, \frac{\pi(x)}{\Pi_1} \bar{t}_{x1} \\
&= \int_{-\infty}^{-a} dx \, \frac{\pi(x)}{\Pi_1} t_{x-a} + \bar{t}_{-a2} - \frac{1}{\Pi_1} \left[\zeta - \int_{-a}^{+\infty} dx \, \pi(x) \bar{t}_{x1} \right] \\
&= \int_{-\infty}^{-a} dx \, \frac{\pi(x)}{\Pi_1} t_{x-a} + \bar{t}_{-a2} - \frac{1}{\Pi_1} \left[\zeta^{orig} - \int_{-a}^{+\infty} dx \, \pi(x) (t_{x-a} + \bar{t}_{-a1}) \right] \\
&= \int_{-\infty}^{-a} dx \, \frac{\pi(x)}{\Pi_1} t_{x-a} + \bar{t}_{-a2} - \frac{1}{\Pi_1} \left[\int_{-\infty}^{-a} dx \, \pi(x) t_{x-a} - (1 - \Pi_1) \bar{t}_{-a1} \right] \\
&= \bar{t}_{-a2} + \frac{1 - \Pi_1}{\Pi_1} \bar{t}_{-a1}.
\end{aligned}
\tag{A15}
$$

Substituting in Eq. (A13), one finally obtains a relation between MFPTs and the stationary distribution of the coarse-grained that holds when the Kemeny constant is maximized with respect to the two boundary positions $-a$ and a.

$$
\Pi_1 t_{-aa} - \bar{t}_{a2} + \frac{\Pi_2}{\Pi_1} \bar{t}_{-a1} = 0. \tag{A16}
$$

This expression will be tested in Sec. III A 1.

A more interpretable result can be derived as follows: Defining

$$
\hat{t}_{J\alpha} = \int_J dx \, \frac{\pi(x)}{\Pi_J} (t_{x\alpha} - \bar{t}_{xJ}) \tag{A17}
$$

and using Eqs. (33) and (A11), we can interpret $\hat{t}_{1-a} = -\bar{t}_{-a2} + t_{12}^{CG}$ as the MFPT to cluster 1 from its right boundary. Then, rearranging Eq. (A10), we have

$$
\begin{aligned}
\frac{\partial}{\partial a} \zeta^{CG}/2 &= -\pi(-a)(\bar{t}_{-a2} - \bar{t}_{-a1}) + \pi(a) \int_{-\infty}^{-a} dx \, p_2(x)(t_{xa} - \bar{t}_{x2}) \\
&+ \pi(-a) \int_{-\infty}^{-a} dx \, p_2(x)(t_{x-a} - \bar{t}_{x2}) + \pi(-a)(t_{12}^{CG} - \bar{t}_{-a2}).
\end{aligned}
\tag{A18}
$$

Using the symmetry of the potential, we can expand the above equation as follows (this corresponds to starting the derivation from $\zeta^{CG} = \Pi_1 t_{12}^{CG} + \Pi_3 t_{32}^{CG}$ instead of $\zeta^{CG} = 2\Pi_1 t_{12}^{CG}$):

$$
\begin{aligned}
\frac{\partial}{\partial a} \zeta^{CG} &= -\pi(-a)(\bar{t}_{-a2} - \bar{t}_{-a1}) + \pi(a) \int_{-\infty}^{-a} dx \, p_2(x)(t_{xa} - \bar{t}_{x2}) + \pi(-a) \int_{-\infty}^{-a} dx \, p_2(x)(t_{x-a} - \bar{t}_{x2}) \\
&+ \pi(-a)(t_{12}^{CG} - \bar{t}_{-a2}) - \pi(a)(\bar{t}_{a2} - \bar{t}_{a3}) + \pi(a) \int_{a}^{+\infty} dx \, p_2(x)(t_{xa} - \bar{t}_{x2}) \\
&+ \pi(-a) \int_{a}^{+\infty} dx \, p_2(x)(t_{x-a} - \bar{t}_{x2}) + \pi(a)(t_{32}^{CG} - \bar{t}_{a2}).
\end{aligned}
\tag{A19}
$$

Writing

$$
\begin{aligned}
\frac{\partial \Pi_2 t_{22}^{CG}}{\partial a} &= \pi(a)(\bar{t}_{a2} - \bar{t}_{a2}) - \pi(-a)(\bar{t}_{-a2} - \bar{t}_{-a2}) + \int_{-a}^{a} dx \, \pi(x) \left[\frac{\partial \bar{t}_{x2}}{\partial a} - \frac{\partial \bar{t}_{x2}}{\partial a} \right] \\
&= \int_{-a}^{a} dx \, \pi(x) \{ p_2(a) t_{xa} + p_2(-a) t_{x-a} - [p_2(a) + p_2(-a)] \bar{t}_{x2} - p_2(a) t_{xa} - p_2(-a) t_{x-a} + [p_2(a) + p_2(-a)] \bar{t}_{x2} \} \\
\\
&= \int_{-a}^{a} dx \, \pi(x) p_2(a)(t_{xa} - \bar{t}_{x2} - t_{xa}) + \int_{-a}^{a} dx \, \pi(x) p_2(-a)(t_{x-a} - \bar{t}_{x2} - t_{x-a}) + \int_{-a}^{a} dx \, \pi(x) [p_2(a) + p_2(-a)] \bar{t}_{x2} \\
&= \int_{-a}^{a} dx \, \pi(a) p_2(x)(t_{xa} - \bar{t}_{x2}) + \int_{-a}^{a} dx \, \pi(-a) p_2(x)(t_{x-a} - \bar{t}_{x2}) + \pi(a) \int_{-a}^{a} dx \, p_2(x)(\bar{t}_{x2} - t_{xa}) + \pi(-a) \int_{-a}^{a} dx \, p_2(x)(\bar{t}_{x2} - t_{x-a}) \\
&= -\pi(a) \hat{t}_{2a} - \pi(-a) \hat{t}_{2-a} + \pi(a) \int_{-a}^{a} dx \, p_2(x)(t_{xa} - \bar{t}_{x2}) + \pi(-a) \int_{-a}^{a} dx \, p_2(x)(t_{x-a} - \bar{t}_{x2}).
\end{aligned}
\tag{A20}
$$

We now add Eq. (A20) to both sides of Eq. (A19). Equation (A20) is equal to zero; hence, it will not alter the result, but will allow to simplify the RHS,

$$
\frac{\partial}{\partial a}\zeta^{CG} = \pi(a)\big[(\hat{t}_{3a}+\bar{t}_{a3})-(\hat{t}_{2a}+\bar{t}_{a2})\big]
$$
$$
+ \pi(-a)\big[(\hat{t}_{1-a}+\bar{t}_{-a1})-(\hat{t}_{2-a}+\bar{t}_{-a2})\big]
$$
$$
+ \pi(a)\int_{-\infty}^{+\infty}dx\,p_2(x)(t_{xa}-\bar{t}_{x2})
$$
$$
+ \pi(-a)\int_{-\infty}^{+\infty}dx\,p_2(x)(t_{x-a}-\bar{t}_{x2}). \tag{A21}
$$

Writing again $\pi(a)p_2(x)=p_2(a)\pi(x)$ in the last two terms, we note that these vanish as $\int_{-\infty}^{+\infty}dx\,\pi(x)t_{x-a}$ and $\int_{-\infty}^{+\infty}dx\,\pi(x)t_{xa}$ are equal to the Kemeny constant of the original system, ζ^{orig}, and $\int_{-\infty}^{+\infty}dx\,\pi(x)\bar{t}_{x2}=\int dy\,p_2(y)\int_{-\infty}^{+\infty}dx\,\pi(x)t_{xy}=\zeta^{\mathrm{orig}}$ as well. Hence, we are left with

$$
\frac{\partial}{\partial a}\zeta^{CG} = \pi(-a)\big[(\hat{t}_{1-a}+\bar{t}_{-a1})-(\hat{t}_{2-a}+\bar{t}_{-a2})\big]
$$
$$
+ \pi(a)\big[(\hat{t}_{3a}+\bar{t}_{a3})-(\hat{t}_{2a}+\bar{t}_{a2})\big]. \tag{A22}
$$

Finally, using the definition of round-trip times given in Eq. (38), we find

$$
\frac{\partial\zeta^{CG}}{\partial a} = \pi(a)\big[t_{3a}^{RT}-t_{2a}^{RT}\big]+\pi(-a)\big[t_{1-a}^{RT}-t_{2-a}^{RT}\big].
$$

From symmetry of the potential, the above equation simplifies to

$$
\frac{\partial\zeta^{CG}}{\partial a} = 2\pi(a)\big[t_{3a}^{RT}-t_{2a}^{RT}\big]. \tag{A23}
$$

APPENDIX B: DIFFUSION ON 1D POTENTIAL: m-STATE CLUSTERING

Considering a continuous one-dimensional space clustered into m clusters, each cluster J can be characterized by two boundaries, b_{J-1} and b_J, i.e., one has $J=(b_{J-1},b_J)$ for each state $J=1,\ldots,m$. The first and final (i.e., m-th) clusters are bounded by the lower and upper boundaries b_0 and b_m of the configuration space, which can take arbitrary values, including $\pm\infty$. Using the continuous formulation of Eq. (30),

$$
\zeta^{CG} = \zeta^{\mathrm{orig}}-\sum_{J=1}^{m}\frac{1}{\Pi_J}\int_J dy\int_J dx\,\pi(y)t_{yx}\pi(x), \tag{B1}
$$

and Eq. (36), we can write

$$
\zeta^{CG} = \zeta^{\mathrm{orig}}-\sum_{J=1}^{m}\int_{b_{J-1}}^{b_J}dy\,\pi(y)\bar{t}_{yJ}. \tag{B2}
$$

The Kemeny derivative with respect to a boundary position b_L takes only contributions from $J=L$ or $J=L+1$, giving

$$
\frac{\partial\zeta^{CG}}{\partial b_L} = -\frac{\partial}{\partial b_L}\Big[\int_{b_{L-1}}^{b_L}dy\,\pi(y)\bar{t}_{yL}+\int_{b_L}^{b_{L+1}}dy\,\pi(y)\bar{t}_{y(L+1)}\Big]. \tag{B3}
$$

We apply the Leibniz differentiation rule to the first (α) and second (β) integrals of Eq. (B3) separately,

$$
\frac{\partial}{\partial b_L}\alpha = \pi(b_L)\bar{t}_{b_L L}+\int_{b_{L-1}}^{b_L}dy\,\pi(y)\frac{\partial}{\partial b_L}\bar{t}_{yL},
$$
$$
\frac{\partial}{\partial b_L}\beta = -\pi(b_L)\bar{t}_{b_L(L+1)}+\int_{b_L}^{b_{L+1}}dy\,\pi(y)\frac{\partial}{\partial b_L}\bar{t}_{y(L+1)}. \tag{B4}
$$

Expanding $\bar{t}_{yL}$ and using the Leibniz rule again, we get

$$
\frac{\partial}{\partial b_L}\bar{t}_{yL} = \frac{\pi(b_L)}{\Pi_L}(t_{yb_L}-\bar{t}_{yL}),
$$
$$
\frac{\partial}{\partial b_L}\bar{t}_{y(L+1)} = \frac{\pi(b_L)}{\Pi_{L+1}}(\bar{t}_{y(L+1)}-t_{yb_L}). \tag{B5}
$$

Inserting Eq. (B5) in Eq. (B4) and rearranging, we have

$$
\frac{\partial\alpha}{\partial b_L} = \pi(b_L)\Big[\bar{t}_{b_L L}+\int_{b_{L-1}}^{b_L}dy\,\frac{\pi(y)}{\Pi_L}(t_{yb_L}-\bar{t}_{yL})\Big],
$$
$$
\frac{\partial\beta}{\partial b_L} = -\pi(b_L)\bar{t}_{b_L(L+1)}-\frac{\pi(b_L)}{\Pi_{L+1}}\int_{b_L}^{b_{L+1}}dy\,\pi(y)(t_{yb_L}-\bar{t}_{y(L+1)}). \tag{B6}
$$

Using for the integrals the shorthand notation provided in Eq. (A16), we can write

$$
\frac{\partial}{\partial b_L}\alpha = \pi(b_L)(\bar{t}_{b_L L}+\hat{t}_{Lb_L}),
$$
$$
\frac{\partial}{\partial b_L}\beta = -\pi(b_L)(\bar{t}_{b_L(L+1)}+\hat{t}_{(L+1)b_L}).
$$

Inserting Eq. (B7) in Eq. (B3),

$$
\frac{\partial\zeta^{CG}}{\partial b_L} = \pi(b_L)\big[(\bar{t}_{b_L(L+1)}+\hat{t}_{(L+1)b_L})-(\bar{t}_{b_L L}+\hat{t}_{Lb_L})\big], \tag{B7}
$$

and using the definition of round-trip times given in Eq. (38), we obtain

$$
\frac{\partial\zeta^{CG}}{\partial b_L} = \pi(b_L)\big[t_{b_L(L+1)}^{RT}-t_{b_L L}^{RT}\big]. \tag{B8}
$$

This shows that the derivative vanishes when the following equality is satisfied:

$$
t_{b_L(L+1)}^{RT} = t_{b_L L}^{RT}. \tag{B9}
$$

This provides an intuitive interpretation of our results: the optimal clusters L ad $L+1$ (obtained by maximizing the Kemeny constant) are those which have the same round-trip times to the separating barrier b_L.

APPENDIX C: DISCRETE RANDOM WALK ON 1D POTENTIAL: m-STATE CLUSTERING

Here we consider the case of a 1D discrete system, i.e., a random walk on a linear chain. In discrete systems, we replace the partial derivative of the coarse-grained Kemeny constant ζ^{CG}, with respect to the barrier position b_L, with the finite difference between the assignments of a node b_L to clusters L and $L + 1$, respectively. In particular, if the clusters are ordered so that $J < J + 1 < J + 2 < \cdots$, moving the barrier α to "the right," corresponds to moving the node b_L to "the left," i.e., from cluster $J + 1$ to J. Hence, the equivalent to $\partial \zeta^{CG} / \partial b_L$ in discrete space is

$$\Delta \zeta^{CG}(b_L, L + 1 \to L) = \zeta^{CG}(b_L \in L) - \zeta^{CG}(b_L \in L + 1). \quad (C1)$$

We can obtain a computationally efficient formula for the finite difference of the coarse-grained Kemeny constant. Following a similar derivation as in the previous section, expanding Eq. (B7) and replacing the integrals with sums, we obtain

$$
\begin{aligned}
\frac{\Delta \zeta^{CG}}{\pi_{b_L}} ={}& \frac{1}{\Pi_L^2} \sum_{i \in L} \sum_{j \in L} \pi_j t_{ji} \pi_i - \frac{1}{\Pi_L} \sum_{j \in L} \pi_j t_{jb_L} \\
& - \frac{1}{\Pi_L} \sum_{i \in L} \pi_i t_{b_L i} - \frac{1}{\Pi_{L+1}^2} \sum_{i \in L+1} \sum_{j \in L+1} \pi_i \pi_j t_{ji} \\
& + \frac{1}{\Pi_{L+1}} \sum_{j \in L+1} \pi_j t_{jb_L} + \frac{1}{\Pi_{L+1}} \sum_{i \in L+1} \pi_i t_{b_L i}. \quad (C2)
\end{aligned}
$$

We use the following property of the mean first passage times in one-dimensional space:

$$
t_{ji} = \begin{cases} t_{jb_j} + t_{b_j j} & \text{if} \quad j > b_j > i, \\ t_{jb_j} - t_{ib_j} & \text{if} \quad j < i < b_j, \end{cases} \quad (C3)
$$

and expand the double sums in Eq. (C2) as follows:

$$
\begin{aligned}
\sum_{i \in L} \sum_{j \in L} \pi_j t_{ji} \pi_i &= \sum_{i \in L} \pi_i \left[\sum_j \pi_j t_{ji} - \sum_{j \notin L} \pi_j t_{ji} \right] = \sum_{i \in L} \pi_i \left(\zeta^{orig} - \sum_{S=1}^{L-1} \sum_{j \in S} \pi_j t_{ji} - \sum_{S=L+1}^{M} \sum_{j \in S} \pi_j t_{ji} \right) = \sum_{i \in L} \pi_i \left[\zeta^{orig} - \sum_{S=1}^{L-1} \sum_{j \in S} \pi_j \left(t_{jb_L} - t_{ib_L} \right) \right. \\
&\left. - \sum_{S=L+1}^{M} \sum_{j \in S} \pi_j \left(t_{jb_L} + t_{b_L i} \right) \right] = \Pi_L \zeta^{orig} - \sum_{i \in L} \sum_{S=1}^{L-1} \sum_{j \in S} \pi_i \pi_j \left(t_{jb_L} - t_{ib_L} \right) - \sum_{i \in L} \sum_{S=L+1}^{M} \sum_{j \in S} \pi_i \pi_j \left(t_{jb_L} + t_{b_L i} \right) \\
&= \Pi_L \left(\zeta^{orig} - \sum_{j \notin L} \pi_j t_{jb_L} \right) + \sum_{S=1}^{L-1} \Pi_S \sum_{i \in L} \pi_i t_{ib_L} - \sum_{S=L+1}^{M} \Pi_S \sum_{i \in L} \pi_i t_{b_L i} = \sum_{S=1}^{L} \Pi_S \sum_{j \in L} \pi_j t_{jb_L} - \sum_{S=L+1}^{M} \Pi_S \sum_{i \in L} \pi_i t_{b_L i},
\end{aligned}
$$

and, similarly,

$$
\sum_{i \in L+1} \sum_{j \in L+1} \pi_j t_{ji} \pi_i = \sum_{S=L+1}^{M} \Pi_S \sum_{j \in L+1} \pi_j t_{jb_L} - \sum_{S=1}^{L} \Pi_S \sum_{i \in L+1} \pi_i t_{b_L i}.
$$

Inserting these equations into Eq. (C2) yields the finite difference of the Kemeny constant on a linear chain

$$
\begin{aligned}
\frac{\Delta \zeta^{CG}}{\pi_{b_L}} ={}& \frac{\sum_{S<L} \Pi_S}{\Pi_L^2} \sum_{j \in L} \pi_j t_{jb_L} - \frac{\sum_{S \geq L} \Pi_S}{\Pi_L^2} \sum_{i \in L} \pi_i t_{b_L i} \\
& + \frac{\sum_{S \leq L+1} \Pi_S}{\Pi_{L+1}^2} \sum_{i \in L+1} \pi_i t_{b_L i} - \frac{\sum_{S>L+1} \Pi_S}{\Pi_{L+1}^2} \sum_{j \in L+1} \pi_j t_{jb_L}. \quad (C4)
\end{aligned}
$$

APPENDIX D: RANDOM WALKS ON COMPLEX NETWORKS: m-STATE CLUSTERING

Below we will show that the result obtained in Appendix B holds in complex networks, up to an approximation.

As a discrete equivalent of the Kemeny constant-derivative with respect to barrier position, we consider the finite difference in the coarse-grained Kemeny constant ζ^{CG} between the case when node α is assigned to cluster B or A,

$$\Delta \zeta^{CG}(\alpha, A \to B) = \zeta^{CG}(\alpha \in B) - \zeta^{CG}(\alpha \in A). \quad (D1)$$

For convenience, we define F_{JI} as follows:

$$F_{JI} = \sum_{i \in I} \sum_{j \in J} \pi_j t_{ji} \pi_i. \quad (D2)$$

This leads to a shorter formulation of Eq. (30) and Eq. (34),

$$\zeta^{CG} = \zeta^{orig} - \sum_J \frac{F_{JJ}}{\Pi_J}, \quad (D3)$$

$$t_{JI}^{CG} = \frac{F_{JI}}{\Pi_J \Pi_I} - \frac{F_{JJ}}{\Pi_J^2}. \quad (D4)$$

If we consider splitting a cluster Γ into two clusters A and B, i.e., $\Gamma = A \cup B$, the following property holds:

$$F_{\Gamma\Gamma} = F_{AA} + F_{AB} + F_{BA} + F_{BB}. \quad (D5)$$

Additionally, we define $A^+ = A \cup \{\alpha\}$ and $B^+ = B \cup \{\alpha\}$, i.e., as the union of the node α and the clusters A and B respectively, implying that A and B do not contain α; hence,

$$\Pi_{A^+} = \Pi_A + \pi_\alpha. \quad (D6)$$

19 September 2023 21:32:22

Using Eq. (D3) and dropping $(\alpha, A \to B)$ for simplicity, we rewrite Eq. (D1) as follows:

$$\Delta\zeta^{CG} = \left[\zeta^{orig} - \sum_{J \notin \{A,B\}} \frac{F_{JJ}}{\Pi_J} - \frac{F_{B^+B^+}}{\Pi_{B^+}} - \frac{F_{AA}}{\Pi_A} \right]$$
$$- \left[\zeta^{orig} - \sum_{J \notin \{A,B\}} \frac{F_{JJ}}{\Pi_J} - \frac{F_{A^+A^+}}{\Pi_{A^+}} - \frac{F_{BB}}{\Pi_B} \right]. \quad (D7)$$

Simplifying the first two terms in the brackets and rearranging, we have

$$\Delta\zeta^{CG} = \left[\frac{F_{BB}}{\Pi_B} - \frac{F_{B^+B^+}}{\Pi_{B^+}} \right] - \left[\frac{F_{AA}}{\Pi_A} - \frac{F_{A^+A^+}}{\Pi_{A^+}} \right]$$
$$= \frac{\Pi_{B^+}F_{BB} - \Pi_B F_{B^+B^+}}{\Pi_{B^+}\Pi_B} - \frac{\Pi_{A^+}F_{AA} - \Pi_A F_{A^+A^+}}{\Pi_{A^+}\Pi_A}. \quad (D8)$$

Expanding $F_{X^+X^+}$ and Π_{X^+} on the numerator, by using Eqs. (D5) and (D6), respectively, and simplifying the cancelling terms

$$\Delta\zeta^{CG} = \left[\frac{\Pi_\alpha F_{BB}}{\Pi_{B^+}\Pi_B} - \frac{F_{\alpha B} + F_{B\alpha} + F_{\alpha\alpha}}{\Pi_{B^+}} \right]$$
$$- \left[\frac{\Pi_\alpha F_{AA}}{\Pi_{A^+}\Pi_A} - \frac{F_{\alpha A} + F_{\alpha A} + F_{\alpha\alpha}}{\Pi_{A^+}} \right]. \quad (D9)$$

From $t_{\alpha\alpha} = 0$, we have $F_{\alpha\alpha} = 0$, and then dividing by π_α and rearranging, we have

$$\frac{\Delta\zeta^{CG}}{\pi_\alpha} = \left[\frac{F_{A\alpha}}{\pi_\alpha\Pi_{A^+}} - \frac{F_{AA}}{\Pi_{A^+}\Pi_A} + \frac{F_{\alpha A}}{\pi_\alpha\Pi_{A^+}} \right]$$
$$- \left[\frac{F_{B\alpha}}{\pi_\alpha\Pi_{B^+}} - \frac{F_{BB}}{\Pi_{B^+}\Pi_B} + \frac{F_{\alpha B}}{\pi_\alpha\Pi_{B^+}} \right]. \quad (D10)$$

Under the approximation $\Pi_A^+ \simeq \Pi_A$ and $\Pi_B^+ \simeq \Pi_B$, which hold for $\pi_\alpha \ll \Pi_A, \Pi_B$,

$$\frac{\Delta\zeta^{CG}}{\pi_\alpha} = \left[\frac{F_{A\alpha}}{\pi_\alpha\Pi_A} - \frac{F_{AA}}{\Pi_A\Pi_A} + \frac{F_{\alpha A}}{\pi_\alpha\Pi_A} \right] - \left[\frac{F_{B\alpha}}{\pi_\alpha\Pi_B} - \frac{F_{BB}}{\Pi_B\Pi_B} + \frac{F_{\alpha B}}{\pi_\alpha\Pi_B} \right]. \quad (D11)$$

We can use the formula for the coarse-grained mean first passage times in terms of F, Eq. (D4), and simplify

$$\frac{\Delta\zeta^{CG}}{\pi_\alpha} = \left[t_{A\alpha}^{CG} + t_{\alpha A}^{CG} \right] - \left[t_{B\alpha}^{CG} + t_{\alpha B}^{CG} \right]. \quad (D12)$$

As $t_{A\alpha}^{CG}$ is the mean first passage time from a cluster that contains a single node α to cluster A, this is the discrete-space analog of the mean first passage time $\hat{t}_{A\alpha}$ defined in Eq. (A16) for diffusion in continuous space, and, similarly, $t_{\alpha A}^{CG}$ is analogous to $\bar{t}_{\alpha A}$ defined in Eq. (36).

Upon defining, as before, the round-trip time $t_{\alpha A}^{RT} = t_{A\alpha}^{CG} + t_{\alpha A}^{CG}$, we can finally rewrite (D12) in the intuitive way

$$\Delta\zeta^{CG}(\alpha, A \to B) = \pi_\alpha \left[t_{\alpha A}^{RT} - t_{\alpha B}^{RT} \right]. \quad (D13)$$

This suggests that, given a network with M clusters, the cluster B to which a new node α should be assigned, in order to maximize the Kemeny constant, is the one with minimum round trip time to α, provided that $\pi_\alpha \ll \Pi_L \forall L = 1, \ldots, M$.

REFERENCES

[1] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[2] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[3] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, 2006).
[4] M. Newman, A.-L. Barabasi, and D. Watts, *The Structure and Dynamics of Networks* (Princeton University Press, Princeton, 2006).
[5] M. Newman, *Networks: An Introduction* (Oxford University Press, 2010).
[6] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical Processes on Complex Networks* (Cambridge University Press, 2012).
[7] N. Masuda, M. A. Porter, and R. Lambiotte, Phys. Rep. **716–717**, 1 (2017).
[8] T. Carletti, D. Fanelli, and R. Lambiotte, J. Phys.: Complexity **2**, 015011 (2021).
[9] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, Nature **407**, 651 (2000).
[10] R. Guimera and L. A. N. Amaral, Nature **433**, 895 (2005).
[11] C. Martelli, A. De Martino, E. Marinari, M. Marsili, and I. P. Castillo, Proc. Natl. Acad. Sci. U. S. A. **106**, 2607 (2009).
[12] A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani, Complexus **1**, 38 (2003).
[13] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou *et al.*, Nucleic Acids Res. **43**, D447 (2015).
[14] A. Laddach, J. C.-F. Ng, S. S. Chung, and F. Fraternali, Curr. Opin. Struct. Biol. **50**, 82 (2018).
[15] E. Agliari, A. Barra, A. Galluzzi, F. Guerra, and F. Moauro, Phys. Rev. Lett. **109**, 268101 (2012).
[16] G. Parisi, Proc. Natl. Acad. Sci. U. S. A. **87**, 429 (1990).
[17] K. Anand and R. Kühn, Phys. Rev. E **75**, 016111 (2007).
[18] R. Kutner, M. Ausloos, D. Grech, T. Di Matteo, C. Schinckus, and H. E. Stanley, Physica A **516**, 240 (2019).
[19] M. E. Newman and J. Park, Phys. Rev. E **68**, 036122 (2003).
[20] E. Agliari, R. Burioni, and P. Contucci, "A diffusive strategic dynamics for social systems," J. Stat. Phys. **139**, 478–491 (2017).
[21] M. E. Newman, Phys. Rev. E **64**, 016132 (2001).
[22] R. Guimerà, B. Uzzi, J. Spiro, and L. A. N. Amaral, Science **308**, 697 (2005).
[23] M. E. J. Newman, *The Structure and Dynamics of Networks* (Princeton University Press, 2011), pp. 221–226.
[24] R. Pastor-Satorras and A. Vespignani, Phys. Rev. Lett. **86**, 3200 (2001).
[25] S. Moore and T. Rogers, Phys. Rev. Lett. **124**, 068301 (2020).
[26] M. Koniaris, I. Anagnostopoulos, and Y. Vassiliou, J. Complex Netw. **6**(2), 243–268 (2018).
[27] L. Gamberi, Y.-P. Förster, E. Tzanis, A. Annibale, and P. Vivo, Sci. Rep. **11**, 14452 (2021).
[28] L. M. Naeni, H. Craig, R. Berretta, and P. Moscato, PLoS One **11**, e0157988 (2016).
[29] V. Labatut and X. Bost, ACM Comput. Surv. **52**, 1 (2019).
[30] N. Lempesis, D. G. Tsalikis, G. C. Boulougouris, and D. N. Theodorou, J. Chem. Phys. **135**, 204507 (2011).
[31] M. A. Novotny, Phys. Rev. Lett. **74**, 1 (1995).
[32] M. Stamatakis and D. G. Vlachos, Comput. Chem. Eng. **35**, 2602 (2011).
[33] W. Wang, T. Liang, F. K. Sheong, X. Fan, and X. Huang, J. Chem. Phys. **149**, 072337 (2018).

19 September 2023 21:32:22

[34]D. J. Sharpe and D. J. Wales, J. Chem. Phys. **155**, 140901 (2021).

[35]M. E. J. Newman, Proc. Natl. Acad. Sci. U. S. A. **103**, 8577 (2006).

[36]R. Lambiotte, J.-C. Delvenne, and M. Barahona, IEEE Trans. Network Sci. Eng. **1**, 76 (2014).

[37]D. Aloise, A. Deshpande, P. Hansen, and P. Popat, Mach. Learn. **75**, 245–248 (2009).

[38]S. Fortunato, Phys. Rep. **486**, 75 (2010).

[39]S. Lloyd, IEEE Trans. Inf. Theory **28**, 129 (1982).

[40]R. S. Berry and R. Breitengraser-Kunz, Phys. Rev. Lett. **74**, 3951 (1995).

[41]K. D. Ball, R. S. Berry, R. E. Kunz, F.-Y. Li, A. Proykova, and D. J. Wales, Science **271**, 963 (1996).

[42]R. E. Kunz, P. Blaudeck, K. H. Hoffmann, and R. S. Berry, J. Chem. Phys. **108**, 2576 (1998).

[43]P. D. Kolokathis and D. N. Theodorou, J. Chem. Phys. **137**, 034112 (2012).

[44]G. C. Boulougouris and D. N. Theodorou, J. Chem. Phys. **127**, 084903 (2007).

[45]G. C. Boulougouris and D. Frenkel, J. Chem. Theory Comput. **1**, 389 (2005).

[46]N.-V. Buchete and G. Hummer, J. Phys. Chem. B **112**, 6057 (2008).

[47]G. R. Bowman, V. S. Pande, and F. Noe, *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation* (Springer, 2014); available at https://link.springer.com/book/10.1007/978-94-007-7606-7.

[48]V. S. Pande, K. Beauchamp, and G. R. Bowman, Methods **52**, 99 (2010).

[49]B. E. Husic and V. S. Pande, J. Am. Chem. Soc. **140**, 2386 (2018).

[50]F. Noé and E. Rosta, J. Chem. Phys. **151**, 190401 (2019).

[51]A. Kölzsch, E. Kleyheeg, H. Kruckenberg, M. Kaatz, and B. Blasius, R. Soc. Open Sci. **5**, 180438 (2018).

[52]A. Gupta and B. Dhingra, in *2012 Students Conference on Engineering and Systems* (IEEE, 2012), pp. 1–4.

[53]S. Brin and L. Page, Comput. Networks ISDN Syst. **30**, 107–117 (1998).

[54]J. G. Kemeny, J. L. Snell, and A. W. Knapp, *Denumerable Markov Chains* (Springer, 1976), pp. 79–105.

[55]C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, J. Comput. Phys. **151**, 146 (1999).

[56]P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte, Linear Algebra Appl. **315**, 39 (2000).

[57]M. Weber, "Clustering by using a simplex structure," Technical Report No. 04-03, ZIB, Berlin, 2003.

[58]M. Weber, "Improved Perron cluster analysis," Technical Report No. 03-04, ZIB, Berlin, 2003.

[59]M. Weber, W. Rungsarityotin, and A. Schliep, "Perron cluster analysis and its connection to graph partitioning for noisy data," Technical Report No. 04-39, ZIB, Berlin, 2004.

[60]S. Kube and M. Weber, "Identification of metastabilities in monomolecular conformation kinetics," Technical Report No. 06-01, ZIB, Berlin, 2005.

[61]S. Kube and M. Weber, "Conformation kinetics as a reduced model for transition pathways," Technical Report No. 05-43, ZIB, Berlin, 2005.

[62]P. Deuflhard and M. Weber, Linear Algebra Appl. **398**, 161 (2005).

[63]J. M. Carr and D. J. Wales, J. Phys. Chem. B **112**, 8760 (2008).

[64]J. G. Kemeny, J. L. Snell, and A. W. Knapp, *Denumerable Markov Chains: With a Chapter of Markov Random Fields by David Griffeath* (Springer Science & Business Media, 2012), Vol. 40.

[65]M. Weber and K. Fackeldey, "G-PCCA: Spectral clustering for non-reversible Markov chains," Technical Report No. 15-35, ZIB, Berlin, 2015.

[66]F. Nüske, L. Boninsegna, and C. Clementi, J. Chem. Phys. **151**, 044116 (2019).

[67]D. Crommelin and E. Vanden-Eijnden, Multiscale Model. Simul. **9**, 1588 (2011).

[68]A. K. Hartmann and D. W. Heermann, J. Chem. Phys. **108**, 9550 (1998).

[69]J. G. Kemeny and J. L. Snell, *Finite Markov Chains* (Van Nostrand Comp. Int., New York, 1960).

[70]A. Kells, V. Koskin, E. Rosta, and A. Annibale, J. Chem. Phys. **152**, 104108 (2020).

[71]J. Berkhout and B. F. Heidergott, Oper. Res. **67**, 892 (2019).

[72]J. L. Palacios and G. Markowsky, Appl. Math. Comput. **406**, 126283 (2021).

[73]S. Yilmaz, E. Dudkina, M. Bin, E. Crisostomi, P. Ferraro *et al.*, "Kemeny-based testing for COVID-19," PLOS ONE **15**(11), e0242401 (2020).

[74]L. C. Freeman, Sociometry **40**(1), 35–41 (1977).

[75]A. Bavelas, J. Acoust. Soc. Am. **22**, 725 (1950).

[76]L. Martini, A. Kells, R. Covino, G. Hummer, N.-V. Buchete, and E. Rosta, Phys. Rev. X **7**, 031060 (2017).

[77]A. Kells, Z. É. Mihálka, A. Annibale, and E. Rosta, J. Chem. Phys. **150**, 134107 (2019).

[78]C. Geyer, in *23rd Symposium on the Interface between Computing Science and Statistics* (Interface Foundation North America, Fairfax, 1991), p. 156.

[79]K. Hukushima and K. Nemoto, J. Phys. Soc. Jpn. **65**, 1604 (1996).

[80]E. Marinari and G. Parisi, Europhys. Lett. **19**, 451 (1992).

[81]A. K. Hartmann and H. Rieger, *Optimization Algorithms in Physics* (Wiley-VCH, Weinheim, 2001).

[82]*New Optimization Algorithms in Physics*, edited by A. K. Hartmann and H. Rieger (Wiley-VCH, Weinheim, 2004).

[83]A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin, Neurocomputing **267**, 664 (2017).

[84]D. Gfeller and P. De Los Rios, Phys. Rev. Lett. **99**, 038701 (2007).

[85]C. Gotsman, in *2003 International Conference on Shape Modeling* (IEEE, 2003), pp. 165–171.

[86]M. A. Javed, M. S. Younis, S. Latif, J. Qadir, and A. Baig, J. Network Comput. Appl. **108**, 87 (2018).

[87]H. Almeida, D. Guedes, W. Meira, and M. J. Zaki, in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (Springer, 2011), pp. 44–59.

[88]S. G. Kobourov, S. Pupyrev, and P. Simonetto, in EuroVis (Short Papers), 2014.

[89]D. Chandler, J. Chem. Phys. **68**, 2959 (1978).

[90]D. Spielman, *Combinatorial Scientific Computing* (Citeseer, 2012), p. 18.

[91]A. Clauset, M. E. Newman, and C. Moore, Phys. Rev. E **70**, 066111 (2004).

[92]V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, J. Stat. Mech.: Theory Exp. **2008**, P10008.

[93]T. Aynaud, V. D. Blondel, J. L. Guillaume, and R. Lambiotte, *Graph Partitioning* (John Wiley and Sons, 2013), pp. 315–345.

[94]S. Fortunato and M. Barthélemy, Proc. Natl. Acad. Sci. U. S. A. **104**, 36 (2007).

[95]S. Emmons, S. Kobourov, M. Gallant, and K. Börner, PLoS One **11**, e0159161 (2016).

[96]J.-C. Delvenne, S. N. Yaliraki, and M. Barahona, Proc. Natl. Acad. Sci. U. S. A. **107**, 12755 (2010).

[97]M. Weber, "Meshless methods in confirmation dynamics," Ph.D. thesis, Department of Mathematics and Computer Science, Free University of Berlin 2006.

[98]S. Kube and M. Weber, J. Chem. Phys. **126**(2), 024103 (2007).

[99]S. Röblitz and M. Weber, Adv. Data Anal. Classif. **7**, 147 (2013).

[100]G. Hummer and A. Szabo, J. Phys. Chem. B **119**, 9029 (2015).

[101]D. Kannan, D. J. Sharpe, T. D. Swinburne, and D. J. Wales, J. Chem. Phys. **153**, 244108 (2020).

[102]R. Lambiotte, J.-C. Delvenne, and M. Barahona, arXiv:0812.1770 (2008).

[103]M. Weber and K. Fackeldey, see see https://opus4.kobv.de/opus4-zib/frontdoor/index/index/docId/5550 for "G-{PCCA}: Spectral clustering for non-reversible Markov chains" (2015).

[104]N. D. Conrad, M. Weber, and C. Schütte, Multiscale Model. Simul. **14**, 1319 (2016).

[105]A. Szabo, K. Schulten, and Z. Schulten, J. Chem. Phys. **72**, 4350 (1980).

[106]D. J. Bicout and A. Szabo, J. Chem. Phys. **106**, 10292 (1997).

[107]P. Coolen-Schrijner and E. A. Van Doorn, Probab. Eng. Inf. Sci. **16**, 351 (2002).

[108]P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti, in *2011 11th International Conference on Intelligent Systems Design and Applications* (IEEE, 2011), pp. 88–93.

[109]M. Blatt, S. Wiseman, and E. Domany, Phys. Rev. Lett. **76**, 3251 (1996).

[110]J. Reichardt and S. Bornholdt, Phys. Rev. Lett. **93**, 218701 (2004).

[111]F. Faizi, G. Deligiannidis, and E. Rosta, J. Chem. Theory Comput. **16**, 2124 (2020).

[112]J. J. Moreno, H. G. Katzgraber, and A. K. Hartmann, Int. J. Mod. Phys. C **14**, 285 (2003).

[113]D. J. Earl and M. W. Deem, Phys. Chem. Chem. Phys. **7**, 3910 (2005).

19 September 2023 21:32:22

[114] J. Machta, Phys. Rev. E **80**, 056706 (2009).

[115] A. K. Hartmann, *Big Practical Guide to Computer Simulations* (World Scientific, Singapore, 2015).

[116] R. H. Swendsen and J.-S. Wang, Phys. Rev. Lett. **57**, 2607 (1986).

[117] F. Barahona, J. Phys. A: Math. Gen. **15**, 3241 (1982).

[118] L. Lovász, *Combinatorics, Paul Erdős is Eighty* (János Bolyai Mathematical Society, 1993), Vol. 2, p. 4, ISBN: 9638022736.

[119] S. Northshield and J. L. Palacios, Braz. J. Probab. Stat. **9**(2), 169 (1995).

[120] A. M. Berezhkovskii and A. Szabo, J. Chem. Phys. **150**, 054106 (2019).

[121] H. Qiu and E. R. Hancock, IEEE Trans. Pattern Anal. Mach. Intell. **29**, 1873 (2007).

[122] L. Yen, F. Fouss, C. Decaestecker, P. Francq, and M. Saerens, in *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (Springer, 2007), pp. 1037–1045.

[123] R. Behmo, N. Paragios, and V. Prinet, in *2008 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2008), pp. 1–8.

[124] D. A. Kofke, J. Chem. Phys. **117**, 6911 (2002).

[125] A. J. Ballard and D. J. Wales, J. Chem. Theory Comput. **10**, 5599 (2014).

[126] E. Rosta and G. Hummer, J. Chem. Phys. **131**, 165102 (2009).

[127] E. Rosta and G. Hummer, J. Chem. Phys. **132**, 034102 (2010).

[128] P. W. Holland, K. B. Laskey, and S. Leinhardt, Soc. Networks **5**, 109 (1983).

[129] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. U. S. A. **99**, 7821 (2002).

[130] V. Krebs, https://www.cc.gatech.edu/dimacs10/archive/clustering.shtml, 2004, http://www.orgnet.com/ (unpublished).

[131] M. Newman, Network data, books about us politics, http://www-personal.umich.edu/~mejn/netdata/, 2013.

[132] D. A. Bader, H. Meyerhenke, P. Sanders, and D. Wagner, "Graph partitioning and graph clustering," in *10th DIMACS Implementation Challenge Workshop, Atlanta, GA, 13-14 February 2012* (American Mathematical Society and Center for Discrete Mathematics and Theoretical Computer Science, 2013), Vol. 588.

[133] H. Jung, K.-i. Okazaki, and G. Hummer, J. Chem. Phys. **147**, 152716 (2017).

[134] D. J. Sharpe and D. J. Wales, Phys. Rev. E **104**, 015301 (2021).

[135] T. D. Swinburne, D. Kannan, D. J. Sharpe, and D. J. Wales, J. Chem. Phys. **153**, 134115 (2020).

[136] D. J. Wales, J. Phys. Chem. Lett. **13**, 6349 (2022).

19 September 2023 21:32:22