Generalized Markov stability of network communities

Aurelio Patelli,[1, 2] Andrea Gabrielli,[3, 1] and Giulio Cimini[4, 1]

[1]*Istituto dei Sistemi Complessi (CNR) UoS Dipartimento di Fisica,*
"Sapienza" Università di Roma, 00185 Rome (Italy)
[2]*Service de Physique de l'Etat Condensé, UMR 3680 CEA-CNRS,*
Université Paris-Saclay, CEA-Saclay, 91191 Gif-sur-Yvette (France)
[3]*Dipartimento di Ingegneria, Università Roma 3, 00146 Rome (Italy)*
[4]*Dipartimento di Fisica, Università di Roma Tor Vergata, 00133 Rome (Italy)*

We address the problem of community detection in networks by introducing a general definition of Markov stability, based on the difference between the probability fluxes of a Markov chain on the network at different time scales. The specific implementation of the quality function and the resulting optimal community structure thus become dependent both on the type of Markov process and on the specific Markov times considered. For instance, if we use a natural Markov chain dynamics and discount its stationary distribution – that is, we take as *reference process* the dynamics at infinite time – we obtain the standard formulation of the Markov stability. Notably, the possibility to use finite-time transition probabilities to define the reference process naturally allows detecting communities at different resolutions, without the need to consider a continuous-time Markov chain in the small time limit. The main advantage of our general formulation of Markov stability based on dynamical flows is that we work with lumped Markov chains on network partitions, having the same stationary distribution of the original process. In this way the form of the quality function becomes invariant under partitioning, leading to a self-consistent definition of community structures at different aggregation scales.

I. INTRODUCTION

Networks are systems made up of entities (nodes) embedded in a complex pattern of interconnections (links), which occur in a large variety of contexts – ranging from socio-economic systems and infrastructures to biological processes and ecosystems [1–4]. Networks observed in nature have a recurrent set of characteristics, such as fat-tail behavior of the degree distribution, small-world topology, and community structure – the latter referring to the internal organization of nodes into densely connected groups. Identifying the communities of a network means uncovering its mesoscopic structure, and is still an outstanding challenge for network science [5–7].

The first method proposed in the literature to partition a network in communities is based on the maximization of a quality function, the *modularity*, which compares the actual number of links in the network falling inside each community to the expectation of such number under a null network model [8]. The modularity function has been then generalized to various setups, like directed, weighted or bipartite networks (see *e.g.* [9, 10]), and still nowadays represents the benchmark method for community detection [6]. However, by relying on a global null model, the modularity suffers from a resolution limit, that is, it cannot find communities smaller than a minimum size – which depends on the scale of the whole system [11]. Multi-resolution versions of the modularity address this issue using a tunable resolution parameter [12, 13], whereas, the modularity-density functional employs a penalty function for splitting partitions [14]. Another popular approach to community detection consists in fitting the network to a stochastic blockmodel, namely a random graph with built-in communities [15], yet this

approach was recently shown to be equivalent to modularity maximization [16]. Other well known community detection methods use clique percolation [17], spectral graph properties [18], spin glass models [13, 19, 20] or combinatorial arguments – notably this latter method, *Surprise* [21, 22], is nearly unaffected by the resolution limit, but has the opposite drawback of overestimating the number of communities [23].

Another popular branch of community detection methods is based on *random walks* [24]. The idea is that communities correspond to network regions where the walker's dynamics spends a relatively long time, because of the high density of links within communities and the sparse connections across communities. This phenomenon leads to the definition of a quality function known as *Markov stability* [25]. Notably, Markov stability allows interpolating between modularity and spectral clustering by simply varying the time scale of the dynamics [26]. Indeed, such a time scale effectively acts as a resolution parameter, with short scales leading to many small communities and long scales to a few large communities [25, 27]. Using continuous-time random walks in the small time limit can even overcome the resolution limit of the modularity [25]. Among related methods, the *Walktrap* algorithm has been one of the first to use random walks for inferring similarities between nodes whence the network community structure [28]. The popular *Infomap* algorithm instead puts the community detection problem in information-theoretical terms [29, 30]: the functional to be optimized with respect to the network partition is the description length for the moves of a random walker on the network. Hence the codebook and the codewords are based on the transition probabilities and stationary distribution of the random walk. Related

to this, methods based on Boltzmann minimum description length have recently been proposed [31]. Random walks have also been used to partition the links (rather than the nodes) of the network, and thus to uncover community structures using the concept of the line graph [32, 33].

The plethora of community detection methods give similar but not identical results, and indeed no algorithm seems to be optimal for all possible community detection tasks [34, 35]. This happens because community detection is an ill-defined problem: there is no universal definition of communities, and thus no clear guidelines on how to build and assess a community detection method [6, 7]. For instance, the approaches based on the network topology (modularity and blockmodel) or on link combinatorics (surprise) use a null network model to assess the statistical significance of a network partition, and the freedom in choosing the null model introduces a degeneracy in the definition of the community structure. Physics-inspired methods suffer from the same pathology, since changing the definition of the interaction between nodes and the strength of the noise give different phases, whereas, methods based on random walks find different partitions depending on the particular dynamics implemented on the network [36].

Given that the quest for the best method to detect the "true" communities of any network is possibly vain, here we follow up on the complementary viewpoint of random walks methods that any given dynamical process on the network induces a different community structure [26]. We thus consider a general Markov diffusion process on the network, and derive a general quality function for the optimization problem using the transition probability fluxes of the dynamics at different time scales. In this way we generalize previous definitions of the Markov stability, which compare the Markov dynamics at finite times to a reference process given by its stationary distribution (i.e., the dynamics at infinite time) [25, 26]. Indeed by varying the time scales of the Markov dynamics and of the reference process we can detect communities at both higher and lower resolutions. Remarkably, our approach is grounded on the definition of lumped Markov chains on network partitions [37], whose stationary distributions follow the same aggregating rules of the dynamics. Thanks to this property the form of the quality function becomes invariant under network partitioning, leading to a self-consistent definition of communities at different aggregation scales. This leads not only to an elegant theoretical formulation of the problem but also to a convenient recursive algorithm for the optimization of the quality function.

II. MARKOV CHAIN ON NETWORKS

We start by recalling basic definitions and properties of Markov chains on networks. We then introduce lumped Markov chains on network partitions, and illustrate these concepts in the simple case of the natural Markov chain (i.e., the random walk).

A network is a set $\mathcal{N}$ of N nodes, whose pattern of interconnections is described by the adjacency matrix – with generic element A_{ij} giving the weight of the link from node i to node j (in the case of binary networks, $A_{ij} = 1$ if the link $i \to j$ exists and 0 otherwise). A Markov chain on a network is a discrete-time stochastic process that defines a temporal sequence of nodes (the possible states of the chain), and that satisfies the Markov property: the probability to be in any state at a given time step depends only on the state attained at the previous step. The process is thus described by the set of probabilities $\{p_{ij}\}_{i,j \in \mathcal{N}}$ of jumping from node i to node j at a given time step [38].

A Markov chain is *ergodic* if it is non-periodic and in the long time regime it visits each node of the network with a non-zero frequency, which converges to a stationary distribution $\{\pi_i\}_{i \in \mathcal{N}}$ satisfying the eigenvalue relation $\pi_j = \sum_{i \in \mathcal{N}} \pi_i p_{ij}$. The transition probability from node i to node j in a finite number n of jumps, p_{ij}^n, is obtained from the n-th power of the single jump transition probability matrix. Similarly, the expected proportion of times that a chain starting from node i visits node j in the first n jumps is $q_{ij}^n = n^{-1} \sum_{m=1}^{n} p_{ij}^m$. Because of ergodicity, in the long-time limit both these quantities converge to the stationary frequency of visiting node j (which is independent on the initial node i):

$$\lim_{n \to \infty} p_{ij}^n = \lim_{n \to \infty} q_{ij}^n = \omega_{ij} \equiv \pi_j, \qquad (1)$$

where ω_{ij} is the *infinite time* transition probability. Note however that the convergence of p_{ij}^n is exponential with n, whereas, that of q_{ij}^m is algebraic in m. Finally, the stationary *probability flux* from node i to node j is the probability that the chain actually jumps from i to j, and thus is given by the asymptotic joint probability of being in i and successively jump to j:

$$\mathrm{F}_p(i \to j) = \pi_i p_{ij}. \qquad (2)$$

Lumped Markov chain on network partitions

A partition of the network nodes into a set of communities $\{\mathcal{C}\}$ induces an aggregated dynamical process, described by transition probabilities $\{\tilde{p}_{\mathcal{C}\mathcal{C}'}\}$ between communities, that is a function of the original Markov chain. Such aggregated process is not necessarily Markovian, since the new transition probabilities could in principle depend on the whole sequence of visited nodes. However, if the original Markov chain is ergodic, it is possible to define a lumped Markov process by preserving the probability fluxes between communities [37]:

$$\mathrm{F}_{\tilde{p}}(\mathcal{C} \to \mathcal{C}') = \sum_{i \in \mathcal{C}} \sum_{j \in \mathcal{C}'} \mathrm{F}_p(i \to j). \qquad (3)$$

This property is called weak lumpability, and translates into a transition probability from $\mathcal{C}$ to $\mathcal{C}'$ of the form

$$\tilde{p}_{\mathcal{C}\mathcal{C}'} = \frac{\sum_{i\in\mathcal{C}}\sum_{j\in\mathcal{C}'}\pi_i p_{ij}}{\tilde{\pi}_{\mathcal{C}}}, \qquad (4)$$

where $\tilde{\pi}_{\mathcal{C}} = \sum_{i\in\mathcal{C}}\pi_i$. The generic diagonal term of this matrix, $\tilde{p}_{\mathcal{C}\mathcal{C}}$, is the *persistence probability* of community $\mathcal{C}$ [37, 39]. Analogously, we can build the finite and infinite time transition probabilities of the lumped process as $\tilde{p}_{\mathcal{C}\mathcal{C}'}^n = \tilde{\pi}_{\mathcal{C}}^{-1}\sum_{i\in\mathcal{C}}\sum_{j\in\mathcal{C}'}\pi_i p_{ij}^n$, $\tilde{q}_{\mathcal{C}\mathcal{C}'}^n = \tilde{\pi}_{\mathcal{C}}^{-1}\sum_{i\in\mathcal{C}}\sum_{j\in\mathcal{C}'}\pi_i q_{ij}^n$ and $\tilde{\omega}_{\mathcal{C}\mathcal{C}'} = \tilde{\pi}_{\mathcal{C}}^{-1}\sum_{i\in\mathcal{C}}\sum_{j\in\mathcal{C}'}\pi_i\omega_{ij}$ [40].

The natural Markov chain

The natural Markov chain gives the simplest instance of transition probabilities between nodes in a network: the probability p_{ij} to jump from node i to one of its neighbor j is uniform across the neighbors, and zero for unconnected nodes. If the network is weighted, p_{ij} simply becomes proportional to A_{ij}. Hence in general $p_{ij} = A_{ij}/d_i$ where $d_i = \sum_{j\in\mathcal{N}}A_{ij}$ denotes the total weight of outgoing connections for node i.

In the case of undirected networks ($A_{ij} = A_{ji}$ for each i, j) with a single connected component, the natural Markov chain is ergodic and the stationary distribution has the analytic form $\pi_i = d_i/(2L)$, where $2L = \sum_{j\in\mathcal{N}}d_j$. Besides, the chain is *reversible* since it satisfies the detailed balance: the probability fluxes between any two nodes are equal, $\mathrm{F}_p(i\to j) \equiv \pi_i p_{ij} = \pi_j p_{ji} \equiv \mathrm{F}_p(j\to i)$. This relation holds simply because fluxes are proportional to the elements of the adjacency matrix, $\mathrm{F}_p(i\to j) \sim A_{ij}$.

Due to this property, the lumped process of a natural Markov chain on a network partition can be mapped to a new weighted adjacency matrix, whose terms are given by the sum of elements of the original adjacency matrix corresponding to nodes in the considered partitions:

$$\mathrm{F}_{\tilde{p}}(\mathcal{C}\to\mathcal{C}') \sim \tilde{A}_{\mathcal{C}\mathcal{C}'} = \sum_{i\in\mathcal{C}}\sum_{j\in\mathcal{C}'}A_{ij}. \qquad (5)$$

III. COMMUNITY DETECTION WITH LUMPED MARKOV CHAINS

We now use the general dynamical framework of lumped Markov chain introduced above to define a quality function for community detection tasks. We start from two key assumptions on which we base our definition of communities.

Firstly, as stated above, different Markov dynamics induce different partitions of the network. According to the principle behind the Markov stability, communities are regions of the network where the Markov chain remains confined for relatively long time – where "relatively

long" has to be assessed using a *reference process*. The most natural choice is to use a reference that brings *zero information* both on the details of the network topology and on the initial state of the dynamics. The infinite time transition probabilities of eq. (3) satisfy this requirement, but this is just a possible choice – we can as well use as reference the Markov dynamics at any finite time.

Secondly, we require that any community has to be resilient to changes occurring locally elsewhere in the network, or equivalently that a community is a community almost independently on the topological details of the rest of the network. This assumption allows simplifying the assessment of an individual community using a lumped Markov chain with two states: the community itself and the rest of the network. Notably such a two-states Markov process can be described using only the stationary distribution $\tilde{\pi}_{\mathcal{C}}$ and persistence probability $\tilde{p}_{\mathcal{C}\mathcal{C}}$ of the community concerned, since the conservation of probability fluxes implies that the flux from $\mathcal{C}$ to anywhere else is equal to the flux from anywhere else to $\mathcal{C}$ [41].

Generalized Markov stability (GMS)

To find a good network partition, we aim at "maximizing the difference" between the Markov dynamics and the reference process. We can thus build a quality function based on the probability flux difference between these two processes. For simplicity we start by considering the single jump Markov dynamics, using as reference process its asymptotic behavior given by the infinite time transition probabilities. For any two nodes i, j in the original Markov chain we define

$$D_{ij} = \mathrm{F}_p(i\to j) - \mathrm{F}_\omega(i\to j) = \pi_i(p_{ij} - \omega_{ij}), \qquad (6)$$

while for two communities $\mathcal{C}, \mathcal{C}'$ in the lumped chain

$$D_{\mathcal{C}\mathcal{C}'} = \mathrm{F}_{\tilde{p}}(\mathcal{C}\to\mathcal{C}') - \mathrm{F}_{\tilde{\omega}}(\mathcal{C}\to\mathcal{C}') = \tilde{\pi}_{\mathcal{C}}\left(\tilde{p}_{\mathcal{C}\mathcal{C}'} - \tilde{\omega}_{\mathcal{C}\mathcal{C}'}\right) \qquad (7)$$

meaning that $D_{\mathcal{C}\mathcal{C}'} \equiv \sum_{i\in\mathcal{C}}\sum_{j\in\mathcal{C}'}D_{ij}$. From this definition we see that the flux difference internal to community $\mathcal{C}$, $D_{\mathcal{C}\mathcal{C}}$, satisfies the requirement of depending only on quantities related to $\mathcal{C}$ itself – with respect to the rest of the network. As global quality function to assess the quality of a network partition we can thus take the trace

$$\mathcal{M}^{[1,\infty]}(\{\mathcal{C}\}) = \sum_{\mathcal{C}}\tilde{\pi}_{\mathcal{C}}\left(\tilde{p}_{\mathcal{C}\mathcal{C}} - \tilde{\omega}_{\mathcal{C}\mathcal{C}}\right), \qquad (8)$$

representing the probability flux that a random walker remains in a community within one time step, discounting the stationary distribution of the process. More generally, if we consider transition probabilities of n jumps against visiting frequencies of m jumps (with $n < m$) we can define

$$\mathcal{M}^{[n,m]}(\{\mathcal{C}\}) = \sum_{\mathcal{C}}\tilde{\pi}_c\left(\tilde{p}_{\mathcal{C}\mathcal{C}}^n - \tilde{p}_{\mathcal{C}\mathcal{C}}^m\right). \qquad (9)$$

This quality function is a *generalized Markov stability* (GMS). Indeed for $m \to \infty$ the reference process is given by the infinite time transition probability as in eq. (8), which converges to the stationary distribution of the dynamics, and in this case $\mathcal{M}^{[n,\infty]}(\{\mathcal{C}\})$ coincides with the traditional definition of Markov stability [25, 26]. Also, for a natural Markov chain on an undirected network and $n = 1$, $\mathcal{M}^{[1,\infty]}(\{\mathcal{C}\})$ coincides with the standard modularity [9]. This equivalence holds because the modularity relies on a null network model (known as the *Chung-Lu configuration model*) that constrains node degrees [9], and the expectations of link probabilities under this null model coincide with the infinite time transition probabilities of the natural Markov chain.

The Markov stability and its generalized version have conceptual and practical advantages with respect to modularity. First of all, the modularity is based on simple link counts, as well as on a null model for the network topology. Typically, null model implementations are limited to the simple Erdös-Rényi random graph, the configuration model and the (possibly degree-corrected) stochastic blockmodel [16] – the few cases that have an analytic formulation. The Markov stability is instead based on a generic Markov process on the network: besides the natural Markov chain one is free to consider other dynamics, *e.g.*, PageRank [42] or maximal entropic random walks [43], as well as higher-order Markov models [44]. Moreover, GMS is automatically defined in the case of directed networks, at stake with modularity [26, 39].

The peculiar advantage of our generalization of Markov stability is instead the possibility of choosing the reference process, in particular by setting its time horizon. This last aspect in particular relates to the resolution limit of the modularity. According to [5], *"the resolution limit comes from the very definition of modularity, in particular from its random model. The weak point of the random model is the implicit assumption that each vertex can interact with every other vertex, which implies that each part of the network knows about everything else [...] It is certainly more reasonable to assume that each vertex has a limited horizon within the network"*. In terms of Markov stability, this is implemented as in eq. (9) by using a finite time horizon for the reference process. In this way, the dynamics is compared not to its stationary distribution (which is achieved after the walker has explored the whole network), but to the finite-time frequency of visiting nodes (*i.e.*, what the walker is able to explore in a finite number of steps). Thus, it is possible to find smaller communities than with modularity. Note also that previous attempts [25, 26] to overcome the resolution limit with standard Markov stability are based on a continuous-time process in the small time limit, rather than on a different reference process as in eq. (9).

Finally, the definition of the quality function using lumped Markov chains is invariant under hierarchical partitioning of the network. We use this feature when implementing the numerical search of communities (see pseudocode 1) using a variant of the Louvain algorithm [45] and a coarse-graining procedure.

Algorithm 1 Louvain-based algorithm

---

procedure GMS MAXIMIZATION
    input:
    $\{p\} \leftarrow$ transition probabilities between nodes
    $\{\mathcal{C}\}$: initial partition of single-node communities
    list: community membership of each node
    repeat
        $\{\tilde{\mathcal{C}}\} \leftarrow$ MOVES($\{\mathcal{C}\}, \{p\}$)
        if $\mathcal{M}(\{\tilde{\mathcal{C}}\}, \{p\}) > \mathcal{M}(\{\mathcal{C}\}, \{p\})$ **then**
            update *list* according to $\{\tilde{\mathcal{C}}\}$
            $\{p\} \leftarrow$ lumped transition probabilities – eq. (4)
            $\{\mathcal{C}\} \leftarrow$ coarse-grained $\{\tilde{\mathcal{C}}\}$ (one node per community)
        end if
    until $\mathcal{M}$ reaches a maximum
    output *list*
    final step:
    $\{p\} \leftarrow$ transition probabilities between nodes
    $\{\mathcal{C}\}$: partition corresponding to *list*
    $\{\mathcal{C}\} \leftarrow$ MOVES($\{\mathcal{C}\}, \{p\}$)
end procedure

function MOVES($\{\mathcal{C}\}, \{p\}$)
    (repeat a few times)
    for all communities $\mathcal{C} \in \{\mathcal{C}\}$ **do**
        for all nodes $i \in \mathcal{C}$ **do** find $\mathcal{C}' \neq \mathcal{C}$ **such that**
            moving i from $\mathcal{C}$ to $\mathcal{C}'$ maximally increases $\mathcal{M}$
            if such $\mathcal{C}'$ exists **then** move i from $\mathcal{C}$ to $\mathcal{C}'$
            end if
        end for
    end for
end function

Numerical optimization

We first work at the finest level of nodes. We start with a configuration where each node is considered as a different community, giving the corresponding initial value for $\mathcal{M}^{[n,m]}$. The moves we consider are successive changes of community for individual nodes. Each move is accepted if the induced change to $\mathcal{M}^{[n,m]}$ is positive (such variation is computed locally because we consider only moves of single nodes and not of node groups). These moves are repeated until no further increase of $\mathcal{M}^{[n,m]}$ can be achieved.

The communities found through this first procedure are then taken as the meta-nodes of a coarse-grained network, while the Markov process for this new network is defined using the lumpability condition of eq. (4). The local moves described above are then repeated again for this network until a new maximum of $\mathcal{M}^{[n,m]}$ is reached. The corresponding partition is then used to build a more coarse-grained network, and the whole process is repeated until no further moves nor coarse-graining steps can increase $\mathcal{M}^{[n,m]}$.

As a final step, we restart the method from the node level but imposing the community structure just found – that is, we check whether the move of a single node can

refine the optimal partition. This is for instance the case in the Karate Club network [46] (see below), in which a single node switches community because of this last step and the GMS value rises from 0.4188 to 0.4198.

IV. RESULTS

Resolution of generalized Markov stability

We first explore the resolution of $\mathcal{M}^{[n,m]}$ with respect to different choices of Markov times n and m. To this end, we consider a natural Markov chain process on the standard toy network of maximal modularity [11]: a ring-like configuration with N cliques of 5 nodes, each clique being connected to only two other cliques (Figure 1). For this graph with $N = 30$ cliques, standard modularity optimization returns a structure of 15 communities (each composed by a pair of cliques), whereas, for $N = 120$ modularity returns 30 communities (each aggregating four adjacent cliques). Standard Markov stability $\mathcal{M}^{[n,\infty]}$ for $n > 1$ instead finds a community structure that is coarser than what is found by modularity [25]. In particular, since communities are defined as regions of the network where the walker remains confined within n jumps, communities become less in number and bigger in size by increasing the time horizon n of the dynamics – as shown in panel (a) of Figure 1. Notably, if we use a reference process at finite time, we automatically obtain finer communities than with modularity. Panel (b) of Figure 1 shows the case of the simple function $\mathcal{M}^{[1,m]}$: there is a sharp transition for the number of communities at a critical value m^*, below which the true structure of N communities (one for each clique) emerges.

The cliques-graphs considered above is a very simple example, especially because the cliques have the same size. We thus consider an heterogeneous graph of 40 cliques whose size is exponentially distributed (ranging from 5 to 100 nodes). Standard modularity maximization on a realization of this graph returns 33 communities, since it tends to group together the small nearest cliques. Instead $\mathcal{M}^{[1,m]}$ finds the true community structure as soon as $m \lesssim 10^3$. Figure 2 shows a detailed analysis of this configuration. We consider cliques with internal connection probability $\alpha = 1$ on the left and $\alpha = 0.8$ on the right. Again the number of detected communities varies as expected with the parameters n and m that set the resolution of GMS. As in the previous example, by increasing n while keeping m fixed the method finds less communities (it fails to detect the small ones), hence we confirm that n effectively sets the minimum size of detected communities. A more interesting picture is obtained by fixing n while varying m. For $\alpha = 1$ we see the same behavior as that of Figure 1: the method finds the correct number of communities for finite m, whereas, it starts aggregating the smallest cliques when the time horizon of the reference process becomes much larger than the size of the largest clique. Instead for $\alpha = 0.8$
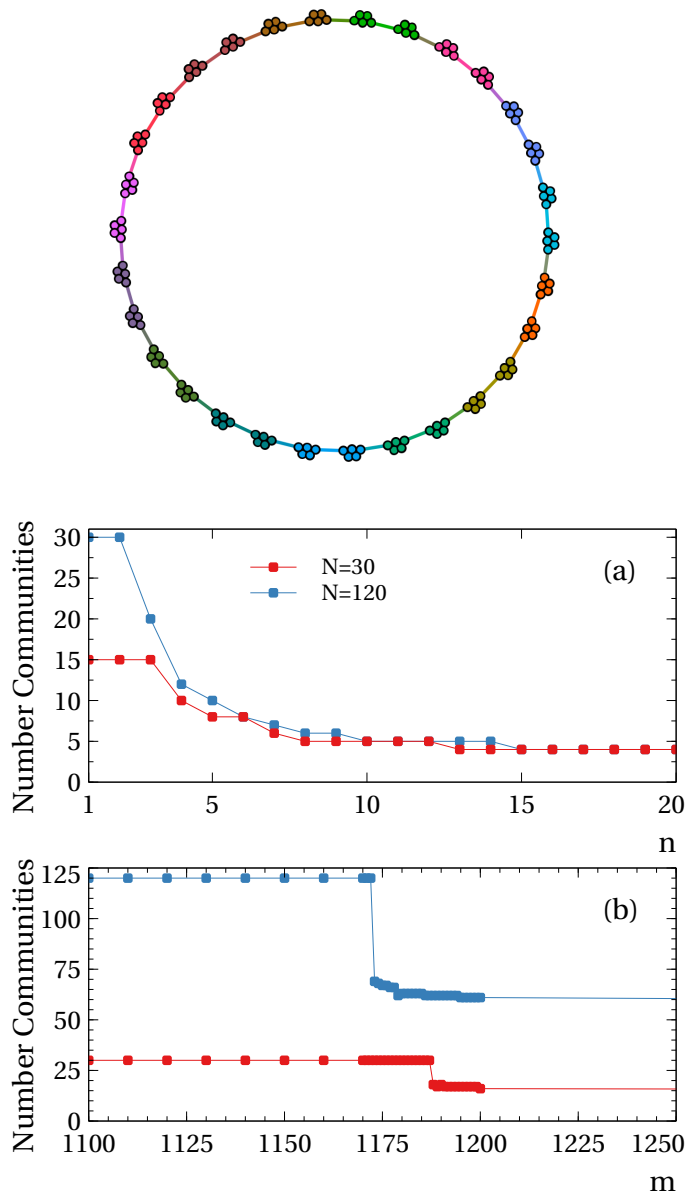

FIG. 1. Resolution of generalized Markov stability for a natural Markov chain on the illustrated network (a ring-like configuration with $N = 30$ cliques of 5 nodes, each clique being connected to only two other cliques). Panel (a): number of communities found by $\mathcal{M}^{[n,\infty]}$ as a function of n. Panel (b): number of communities found by $\mathcal{M}^{[1,m]}$ as a function of m. In both panels we show the cases $N = 30$ (red lines) and $N = 120$ (blue lines).

cliques are not so strongly connected, and fluctuations may induce dense regions internal to cliques. Therefore, we observe a crossover between the region where m is too small to accommodate for the larger cliques (so that the number of detected communities grows with m) and again the regime where m is large enough and the number of detected communities decreases. These examples teach us that we cannot expect to achieve the best per-
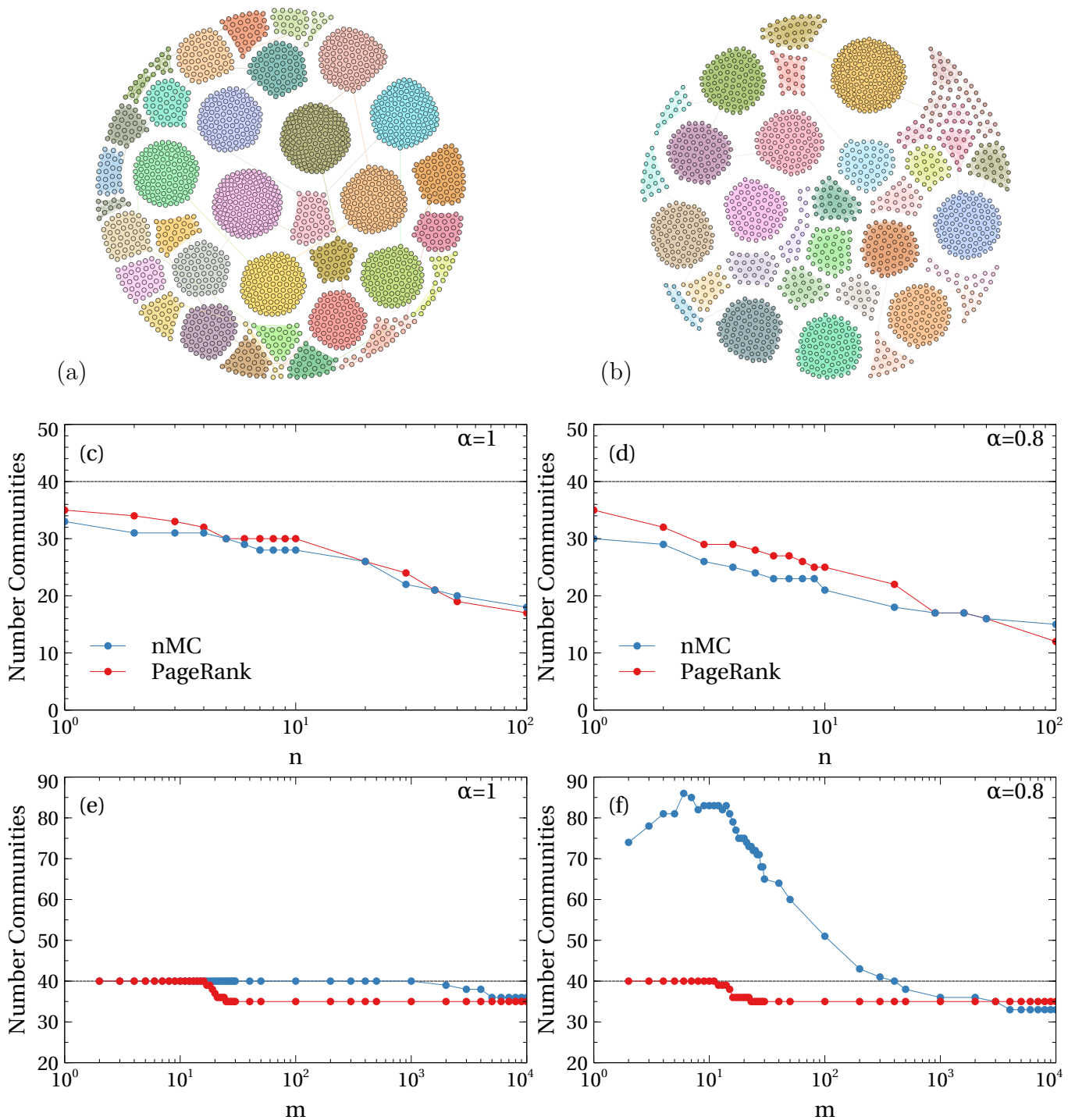
FIG. 2. (a,b) Visual representation of communities found by standard modularity $\mathcal{M}^{[1,\infty]}$ on a ring-like configuration with 40 cliques of varying (exponentially distributed) size, each clique being connected to only two other cliques. The internal connection probability α of the cliques is 1.0 for network (a) and 0.8 for network (b). Each community is represented by a different color. (c,d,e,f) Number of communities identified by generalized Markov stability $\mathcal{M}^{[n,m]}$ as a function of parameters n and m (the time scale of the dynamics and of the reference process) for the same network configuration of panels (a,b). We show GMS implementations using the natural Markov chain (in blue) and PageRank with $\mu/N = 0.15$ (in red). See below for further details on this alternative dynamics.

formance at infinite m, because in this case the method will discard important local information on the network, neither in general at small m for which the horizon of the random walker is simply too limited.

GMS for different random processes

$\mathcal{M}^{[n,m]}$ of eq. (8) is defined for a generic Markov process on the network – the only requirement being the existence of the stationary distribution and of its finite-time version. The induced community structure can thus strongly depend on which process is implemented. Beyond the natural Markov chain, we considered two other processes.

The first one is *PageRank* [42] (see also [39, 47]), which complements the natural Markov chain with a teleportation term allowing for jumps between any two nodes: $p_{ij} = (1 - \mu)A_{ij}/d_i + \mu/N$. In general, teleportation increases the probability to jump outside a community, hence the number of identified communities decreases with the teleportation rate μ. Additionally, PageRank leads to similar results to that if a natural Markov chain with longer time horizon, and at the same time is less sensitive to topological fluctuations (see Figure 2 e,f). Notably, the teleportation rate μ makes the chain ergodic even if the network has disconnected components or if it is directed and has transient parts (that the walker cannot access after leaving them).

The second Markov process we consider is the *maximal entropy random walk* (MERW) [43, 48], also known as the Ruelle-Bowen process in discrete time [26]. MERW transition probabilities are such that all trajectories of given length and given endpoints are equiprobable, and take the form $p_{ij} = (A_{ij}/\lambda)/(\psi_j/\psi_i)$ – where λ is the largest eigenvalue of the adjacency matrix and ψ_i is the i-th component of the normalized eigenvector associated to λ. MERW has strong localization property, imprisoning the walkers in entropic wells [43].

To visually grasp the effect of using a particular Markov dynamics, we show in Figure 3 the communities detected by $\mathcal{M}^{[1,\infty]}$ using natural Markov chain, PageRank and MERW, on the illustrative example of the `Dolphins` network [49] – the network of "swimming together" relations among a group of dolphins. For this network we do not have information on any reference community structure, hence we cannot assess which Markov dynamics performs best. However, despite the identified partitions vary with the Markov process, notably some communities seems more persistent with respect to the specific dynamics employed (in this case, the top left part of the network). Thus comparing the results of multiple dynamics can increase our confidence level on the detected network partition.

GMS versus metadata partitions in real networks

We now put GMS to the test of real and synthetic networks that represent the traditional benchmarks for community detection methods. These networks posses node metadata information that allows defining reference partitions to be used for comparison (see however [35, 50] about the problems of associating metadata groups with
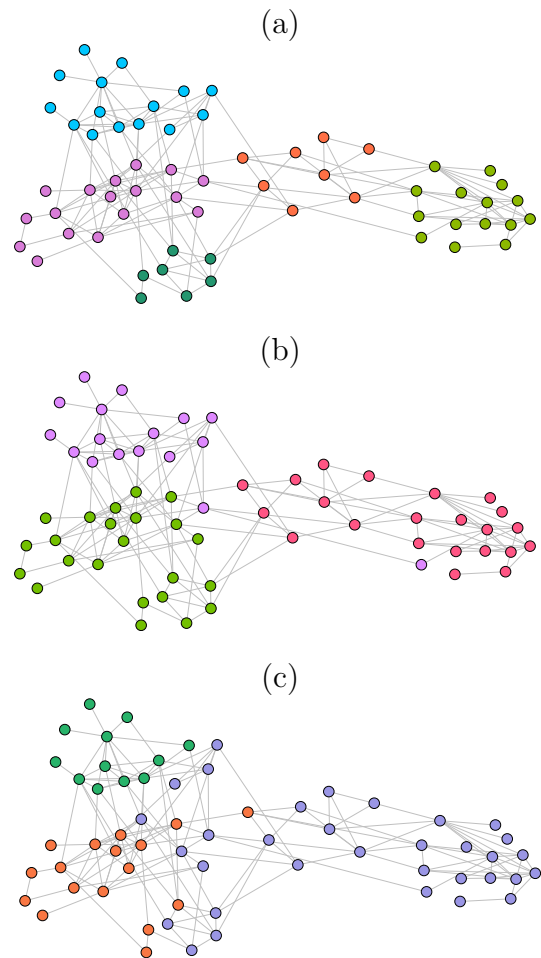

(a)

(b)

(c)

FIG. 3. Communities of the `Dolphins` network found by $\mathcal{M}^{[1,\infty]}$ with natural Markov chain (a), PageRank with $\mu/N = 0.006$ (b) and MERW (c).

topological communities).

We start by briefly describing the datasets we use (that we downloaded from `http://www-personal.umich.edu/~mejn/netdata/`). A full description can be found in the cited references, as well as in [50].

- `football` is the network of American football games between Division IA colleges during season Fall 2000 [8]. Links exist if two teams played any game, and there are 12 groups of teams (conferences) for scheduling intra-group games.

- `karate` is the friendship network of Zachary's karate club [46] that has two natural communities, corresponding to the split of the club in two factions after a dispute between the coach and the treasurer.

- `polblogs` is the network of (undirected) hyperlinks between weblogs on US politics after the 2004 elections [51]. Groups are "liberal" or "conservative" as assigned by either blog directories or self-evaluation.
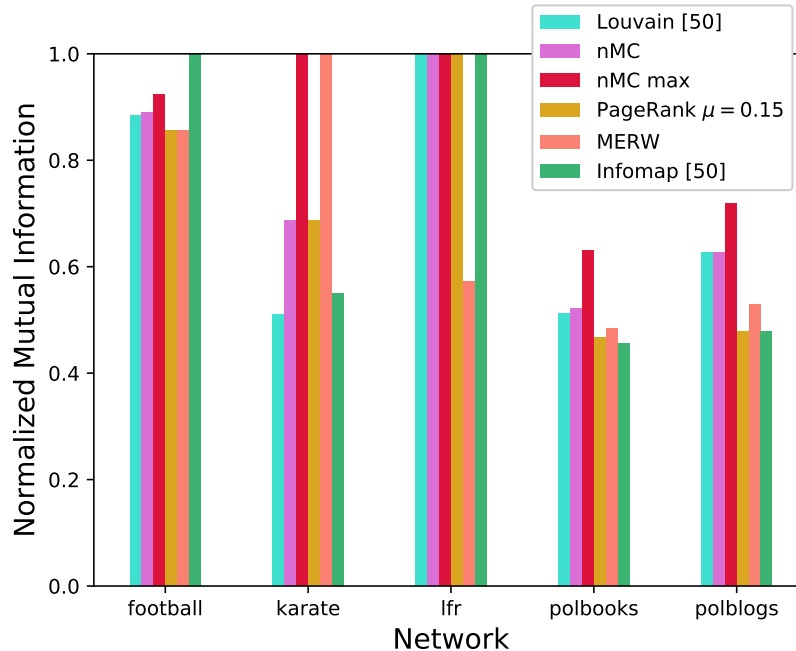
FIG. 4. NMI scores between structural communities found by various methods and metadata groups, for the five networks we consider (scores are clustered by datasets on the horizontal axis). The figure is realized following Figure 5 of [50]. The methods used are as follows: Louvain Modularity; GMS with natural Markov chain $n = 1$ $m = \infty$ (nMC); GMS with natural Markov chain and parameters n and m yielding the highest NMI for that network (nMC max); GMS with PageRank ($\mu/N = 0.15$) $n = 1$ $m = \infty$; GMS with MERW $n = 1$ $m = \infty$; Infomap.

- **polbooks** is the network of books about US politics from 2004 election, taken from Amazon.com [53]. Links represent co-purchasing of books. Groups are based on political alignment: "liberal", "neutral", or "conservative", according to human evaluation.

- Finally, **lfr** is an artificial network with built-in topological communities, generated through the state-of-the-art LFR benchmark [54] (parameters $N = 1000$, 40 small communities 0f size ranging between 10 and 50, and mixing parameter $\frac{1}{2}$). The **lfr** generator code is available at https://sites.google.com/view/santofortunato/software.

We use the Normalized Mutual Information (NMI) [55] to measure the similarity between the network partition induced by a community detection method and the metadata communities of the network. A comparative assessment of how well different methods perform according to this metric is reported in Figure 4. In particular we show the resulting NMI obtained by implementing four different dynamics on GMS: standard natural Markov chain ($n = 1$, $m = \infty$, labeled nMC); natural Markov chain with parameters n and m yielding the highest NMI for that network (nMC max); standard PageRank with $\mu/N = 0.15$ ($n = 1$, $m = \infty$); MERW ($n = 1$, $m = \infty$). We add to the comparison the two state-of-the-art Louvain [45] and Infomap [29] algorithms (for the performance of other methods, we remand the reader to Figure

5 of [50]).

The detailed performance of GMS for varying n and m in shown in Figure 5 separately for each of the considered networks. In the case of **karate**, GMS needs $n > 1$ (but finite) and $m = \infty$ to retrieve a partition corresponding to the two metadata groups. This is an expected outcome because the network is sparse and the two groups are big (compared to the whole network), therefore the random walker cannot fully explore them within just a few steps n. At the same time, m must be large because each community needs to be assessed against the whole network. As side remark, MERW outperforms the other dynamics on most time scales. This happens because MERW is strongly localized on hubs, which in the **karate** network are the coach and the treasurer who are the central members of each group. Moving further, in **lfr** we see on one hand that nMC and PageRank return the metadata groups of the network even at small reference horizon m, because each group is dense but small compared to the network size, and thus does not need to be assessed against the whole network to be retrieved. On the other hand, by increasing n the NMI decreases because the random walker is more likely to travel within groups and thus GMS ends up aggregating the smallest communities. A similar behavior is observed in **football**: a good quality of the partition is obtained at small horizons m, whereas, by increasing n the walker is less likely to stay confined in a group. Finally, at stake with the previous two net-
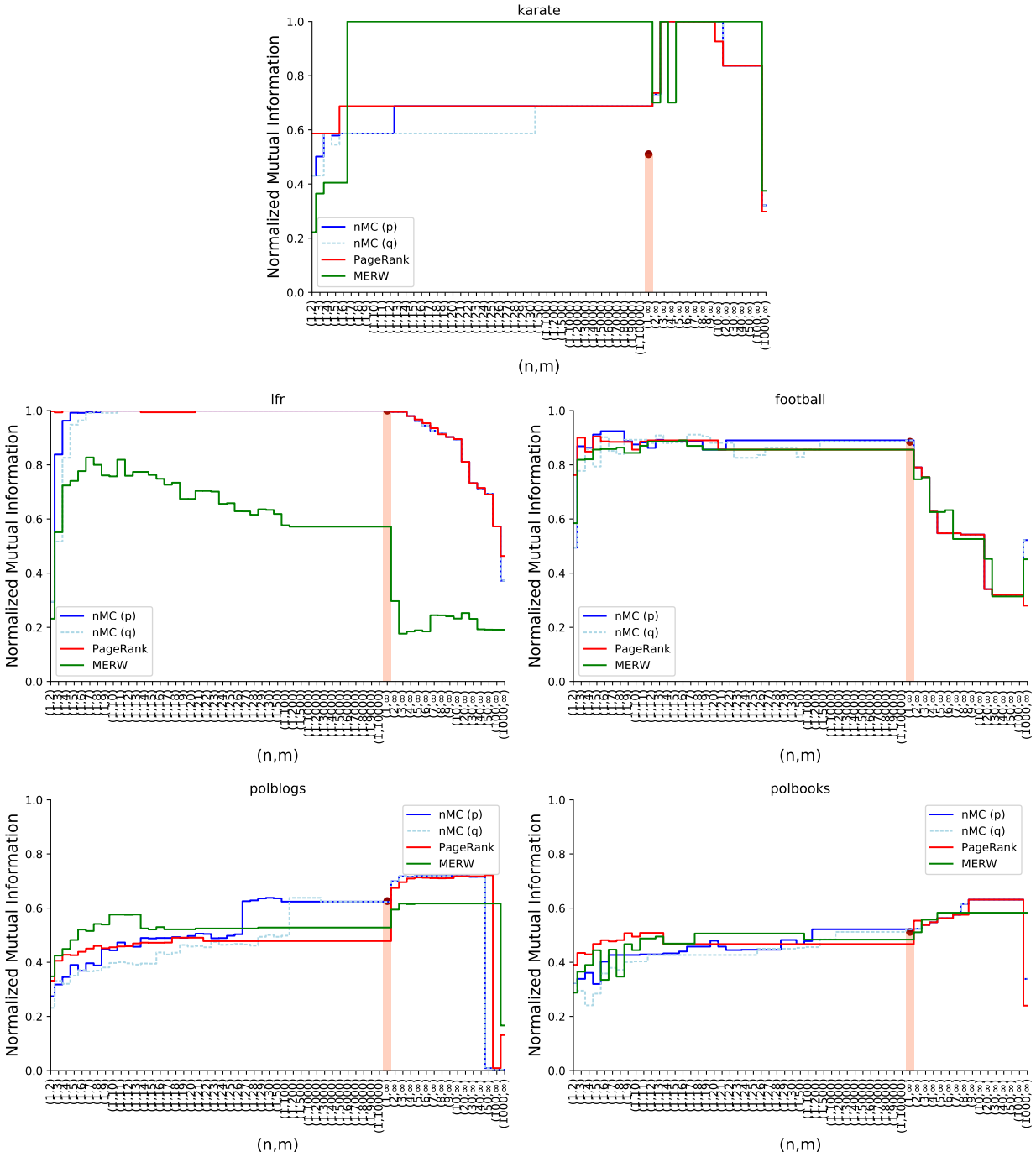
FIG. 5. NMI scores between structural communities found by GMS and metadata communities as a function of the parameters n and m setting the time scale of the dynamics and of the reference process. The dynamics are: natural Markov chain of eq. (9) (nMC (p)), natural Markov chain of eq. (10) (nMC (q)), PageRank with $\mu/N = 0.006$ and MERW. The red dot in correspondence of the vertical bar denotes the NMI value obtained by standard modularity using the Louvain algorithm.

works, the `polblogs` and `polbooks` are sparse and their metadata groups are few (two for `polblogs` and three for `polbooks`) but large. Therefore, similarly to `karate`, these groups are better retrieved at moderately large values of n for which the walker can fully explore each community, and at the same the reference walker must have explored the whole network (m large).

Summarizing, we have found that there is no recipe that performs best in all situations. The optimal performance of GMS as a function of the parameters n and m however provides information on what are the features of the communities that exist in a network. For instance, `lfr` and `football` are both characterized by many small and dense metadata groups, hence GMS work well with small scales of the dynamics (n) and of the reference process (m). `polblogs` and `polbooks` on the contrary have a few large and sparse groups, which are retrieved with wider dynamical horizon n. `karate` belongs to this latter case, but the presence of the two hubs and the consequent degree heterogeneity enhance the performance of MERW.

Alternative reference process

GMS can be as well defined with a reference process measuring the visiting frequencies within m jumps (that is, the q^m matrix) instead of setting a fixed horizon at a temporal scale m (represented by p^m). This leads to a reformulation of eq. (9) as

$$\mathcal{M}^{[n,m]}(\{\mathcal{C}\}) = \sum_{\mathcal{C}} \tilde{\pi}_c \left(\tilde{p}^n_{\mathcal{CC}} - \tilde{q}^m_{\mathcal{CC}} \right). \qquad (10)$$

In this alternative formulation, the reference process contains the contribution of short walks that carry information on the local properties of the network. As explained above, using q instead of p leads to a slower convergence for $m \to \infty$, however the two approaches are qualitatively similar – especially for small values of m (see Figure 5).

Directed networks

As a final remark, we stress that the definition of generalized Markov stability does not depend on the specific network features. Therefore, GMS can be directly implemented on directed networks, provided the considered Markov chain is ergodic (an easy solution for this is the teleportation term of PageRank). Indeed, the case $n = 1$ and $m = \infty$ for simple random walks on directed networks has been studied in [39] as a generalization of standard modularity.

CONCLUSIONS

In this work we reformulated the use of ergodic Markov chains applied to the problem of community detection in networks. Specifically, we defined a lumped Markov process between communities, whose transition probability fluxes are built by aggregating the probability fluxes at the level of nodes. This aggregated process is then used to define a quality function to evaluate a network partition, by requiring the probability fluxes internal to communities (*i.e.*, the persistence probabilities) to be maximally larger than those of a reference case. This results in a generalized version of the Markov stability (GMS).

We remark that the whole theoretical construction of GMS derives from two simple requests: 1) the existence of the reference process, used to assess the persistence probabilities of the dynamics, and 2) the resilience of communities to changes occurring elsewhere in the network, so that the search of communities can be decomposed into multiple two-states problems (for each community, the assessment of the community itself against the rest of the network).

GMS can be implemented with any ergodic Markov dynamics on the network. Additionally, being based on the concept of lumped Markov chains, the GMS quality function is invariant under network partitioning. This means that we can aggregate and disaggregate both nodes and node groups without losing information on the structure and dynamics of the network. This feature is at the basis of the algorithm we developed to optimize the quality function.

Concerning the implementation of GMS, when the reference process corresponds to a dynamics in which all the information on initial conditions and nodes correlations is lost, as in the case of the infinite time transition probability, we obtain the standard formulation of the Markov stability. However considering a reference process with a finite time horizon allows finding communities of varying size – thus overcoming in a natural way the resolution limit typical of the modularity and other approaches. Indeed the time scales of the Markov dynamics and of the reference process effectively set the resolution level of the method. Communities obtained at different resolutions are in general not hierarchical, as in [12]. However, optimizing the GMS quality function with respect to n and m means identifying the size window and other features of the network communities. For a given network structure, the optimal combination of dynamical process, resolution value n and (finite) horizon m [11] can be found *a-posteriori*.

At last we remark that the framework we developed is general and can possibly be applied to other kinds of networks (*e.g.*, bipartite graphs) or to detect overlapping communities. Another interesting research direction would be to compare Markov processes of different nature within the quality function.

[1] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," Reviews of Modern Physics **74**, 47–97 (2002).

[2] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," Physics Reports **424**, 175–308 (2006).

[3] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, "Critical phenomena in complex networks," Reviews of Modern Physics **80**, 1275–1335 (2008).

[4] G. Cimini, T. Squartini, F. Saracco, D. Garlaschelli, A. Gabrielli, and G. Caldarelli, "The statistical physics of real-world networks," Nature Reviews Physics **1**, 58–71 (2019).

[5] S. Fortunato, "Community detection in graphs," Physics Reports **486**, 75–174 (2010).

[6] S. Fortunato and D. Hric, "Community detection in networks: A user guide," Physics Reports **659**, 1–44 (2016).

[7] M. T. Schaub, J.-C. Delvenne, M. Rosvall, and R. Lambiotte, "The many facets of community detection in complex networks," Applied Network Science **2**, 4 (2017).

[8] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," Proceedings of the National Academy of Sciences **99**, 7821–7826 (2002).

[9] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," Physical Review E **69**, 026113 (2003).

[10] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, "Finding statistically significant communities in networks," PLoS ONE **6**, 1–18 (2011).

[11] S. Fortunato and M. Barthélemy, "Resolution limit in community detection," Proceedings of the National Academy of Sciences **104**, 36–41 (2007).

[12] A. Arenas, A. Fernández, and S. Gómez, "Analysis of the structure of complex networks at different resolution levels," New Journal of Physics **10**, 053039 (2008).

[13] J. Reichardt and S. Bornholdt, "Statistical mechanics of community detection," Physical Review E **74**, 016110 (2006).

[14] M. Chen, T. Nguyen, and B. K. Szymanski, "A new metric for quality of network community structure," `http://arxiv.org/abs/1507.04308` (2015).

[15] B. Karrer and M. E. J. Newman, "Stochastic blockmodels and community structure in networks," Physical Review E **83**, 016107 (2011).

[16] M. E. J. Newman, "Equivalence between modularity optimization and maximum likelihood methods for community detection," Physical Review E **94**, 052315 (2016).

[17] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," Nature **435**, 814–818 (2005).

[18] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," Physical Review E **74**, 036104 (2006).

[19] V. A. Traag, P. Van Dooren, and Y. Nesterov, "Narrow scope for resolution-limit-free community detection," Physical Review E **84**, 016114 (2011).

[20] V. A. Traag, G. Krings, and P. Van Dooren, "Significant scales in community structure," Scientific Reports **3**, 2930 (2013).

[21] R. Aldecoa and I. Marín, "Deciphering network community structure by surprise," PLoS ONE **6**, 1–8 (2011).

[22] R. Aldecoa and I. Marín, "Surprise maximization reveals the community structure of complex networks," Scientific Reports **3**, 1060 (2013).

[23] V. A. Traag, R. Aldecoa, and J.-C. Delvenne, "Detecting communities using asymptotical surprise," Physical Review E **92**, 022816 (2015).

[24] N. Masuda, M. A. Porter, and R. Lambiotte, "Random walks and diffusion on networks," Physics Reports **716-717**, 1–58 (2017).

[25] J.-C. Delvenne, S. N. Yaliraki, and M. Barahona, "Stability of graph communities across time scales," Proceedings of the National Academy of Sciences **107**, 12755–12760 (2010).

[26] R. Lambiotte, J.-C. Delvenne, and M. Barahona, "Random walks, Markov processes and the multiscale modular organization of complex networks," IEEE Transactions on Network Science and Engineering **1**, 76–90 (2014).

[27] M. Kheirkhahzadeh, A. Lancichinetti, and M. Rosvall, "Efficient community detection of network flows for varying Markov times and bipartite networks," Physical Review E **93**, 032309 (2016).

[28] P. Pons and M. Latapy, "Computing communities in large networks using random walks," Journal of Graph Algorithms and Applications **10**, 191–218 (2006).

[29] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," Proceedings of the National Academy of Sciences **105**, 1118–1123 (2008).

[30] M. Rosvall, D. Axelsson, and C. T. Bergstrom, "The map equation," The European Physical Journal Special Topics **178**, 13–23 (2009).

[31] J. I. Perotti, C. J. Tessone, A. Clauset, and G. Caldarelli, "Thermodynamics of the Minimum Description Length on community detection," `https://arxiv.org/abs/1806.07005` (2018).

[32] T. S. Evans and R. Lambiotte, "Line graphs, link partitions, and overlapping communities," Physical Review E **80**, 016105 (2009).

[33] A. Gabrielli, R. Mastrandrea, G. Caldarelli, and G. Cimini, "Grand canonical ensemble of weighted networks," Physical Review E **99**, 030301 (2019).

[34] R. Aldecoa and I. Marín, "Exploring the limits of community detection strategies in complex networks," Scientific Reports **3**, 2216 (2013).

[35] L. Peel, D. B. Larremore, and A. Clauset, "The ground truth about metadata and community detection in networks," Science Advances **3**, e1602548 (2017).

[36] V. Zlatic, A. Gabrielli, and G. Caldarelli, "Topologically biased random walk with application for community finding in networks," Physical Review E **82**, 066109 (2010).

[37] C. Piccardi, "Finding and testing network communities by lumped Markov chains," PLoS ONE **6**, e27028 (2011).

[38] We limit our analysis to time-homogeneous Markov chains, for which the transition probabilities do not depend on the current time step.

[39] Y. Kim, S. W. Son, and H. Jeong, "Finding communities in directed networks," Physical Review E **81**, 016103 (2010).

[40] This is generally different from taking the limit $n \to \infty$ of the n jump transition probabilities for the lumped pro-

cess.

[41] An ergodic Markov chain with two possible states (labeled 1 and 2) can be described using only two parameters. Here we take π_1, the stationary distribution of state 1, and p_{12}, the transition probability from state 1 to 2. Firstly, because of normalization we have $\pi_2 = 1 - \pi_1$ and $p_{11} = 1 - p_{12}$. Then, since the chain is reversible we have $F(1 \rightarrow 2) = F(2 \rightarrow 1)$ and thus $p_{21} = p_{12}\pi_1/(1 - \pi_1)$ and $p_{22} = 1 - p_{21}$.

[42] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the Web," World Wide Web Internet and Web Information Systems **54**, 1–17 (1998).

[43] Z. Burda, J. Duda, J. M. Luck, and B. Waclaw, "Localization of the maximal entropy random walk," Physical Review Letters **102**, 160602 (2009).

[44] V. Salnikov, M. T. Schaub, and R. Lambiotte, "Using higher-order markov models to reveal flow-based communities in networks," Scientific Reports **6**, 23194 (2016).

[45] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," Journal of Statistical Mechanics: Theory and Experiment **10**, P10008 (2008).

[46] W. Zachary, "An information flow model for conflict and fission in small groups," Journal of Anthropological Research **33**, 452–473 (1977).

[47] R. Lambiotte and M. Rosvall, "Ranking and clustering of nodes in networks with smart teleportation," Physical Review E **85**, 056107 (2012).

[48] J. K. Ochab and Z. Burda, "Maximal entropy random walk in community detection," The European Physical Journal Special Topics **216**, 73–81 (2013).

[49] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations," Behavioral Ecology and Sociobiology **54**, 396–405 (2003).

[50] D. Hric, R. K. Darst, and S. Fortunato, "Community detection in networks: Structural communities versus ground truth," Physical Review E **90**, 062805 (2014).

[51] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 u.s. election: Divided they blog," in *Proceedings of the 3rd International Workshop on Link Discovery*, LinkKDD '05 (ACM, New York, NY, USA, 2005) pp. 36–43.

[52] The `polblogs` graph features several nodes with no incoming nor outgoing connections, which lead to non-ergodicity for any Markov chain dynamics without teleportation. To avoid this problem we consider only the strongly connected component of the graph, removing 266 nodes over 1490.

[53] V. Krebs, `http://www.orgnet.com/`.

[54] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," Physical Review E **78**, 046110 (2008).

[55] Danon L., A. Díaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," Journal of Statistical Mechanics: Theory and Experiment **2005**, P09008–P09008 (2005).