Identifying Modular Flows on Multilayer Networks Reveals Highly Overlapping Organization in Interconnected Systems

Manlio De Domenico,[1] Andrea Lancichinetti,[2] Alex Arenas,[1] and Martin Rosvall[2]
[1]*Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, 43007 Tarragona, Spain*
[2]*Integrated Science Lab, Department of Physics, Umeå University, SE-901 87 Umeå, Sweden*
(Received 12 September 2014; published 6 March 2015)

To comprehend interconnected systems across the social and natural sciences, researchers have developed many powerful methods to identify functional modules. For example, with interaction data aggregated into a single network layer, flow-based methods have proven useful for identifying modular dynamics in weighted and directed networks that capture constraints on flow processes. However, many interconnected systems consist of agents or components that exhibit multiple layers of interactions, possibly from several different processes. Inevitably, representing this intricate network of networks as a single aggregated network leads to information loss and may obscure the actual organization. Here, we propose a method based on a compression of network flows that can identify modular flows both within and across layers in nonaggregated multilayer networks. Our numerical experiments on synthetic multilayer networks, with some layers originating from the same interaction process, show that the analysis fails in aggregated networks or when treating the layers separately, whereas the multilayer method can accurately identify modules across layers that originate from the same interaction process. We capitalize on our findings and reveal the community structure of two multilayer collaboration networks with topics as layers: scientists affiliated with the Pierre Auger Observatory and scientists publishing works on networks on the arXiv. Compared to conventional aggregated methods, the multilayer method uncovers connected topics and reveals smaller modules with more overlap that better capture the actual organization.

DOI: 10.1103/PhysRevX.5.011027

Subject Areas: Complex Systems, Interdisciplinary Physics

I. POPULAR SUMMARY

The convention to represent different types of interactions in a system with a single type of link no longer is the panacea of network science. Temporal-, memory-, and multiplex-network representations have proven necessary to capture essential structural information in social and biological systems. Many useful tools of conventional network science have quickly been generalized to multilayer networks, but generalizing community-detection algorithms has turned out to be a twofold challenge: What is a community in a multilayer network and how can it be identified? We demonstrate that the information-theoretic and flow-based community-detection method known as the map equation provides an effective answer to both questions. We illustrate the mathematical machinery and demonstrate with an analysis of synthetic and real networks.

II. INTRODUCTION

The multifaceted relationships between numerous components in social and biological systems make them inherently complex to analyze [1,2]. Data about these interactions have become increasingly available, and network analysis has emerged as an essential tool for studying their function [3–5]. For large networks, detailed modeling of individual components and their interactions is unfeasible, and researchers instead seek to simplify and highlight important large-scale functional structures in the networks. Depending on the system under study and the research question at hand, researchers use methods that operate either on the plain topology of the network itself [6,7] or, to capture flow processes through the real system, on dynamics modeled on the network [8,9]. In any case, an important objective is to detect so-called communities [10], topological groups of nodes with higher internal than external density of links compared to null models [11–13] or, alternatively, modules that capture flows for a relatively long time [14–16].

However, community-detection methods generally assume that a single type of static link, weighted and directed at best, can account for all types of interactions between nodes in the network. This assumption oversimplifies the

multifaceted nature of relationships in real systems with important consequences; see Refs. [17,18] for a review. Aggregating multiple types of relationships into a single weighted and directed network can distort both the topology of the network and the dynamics on the network [19]. Take social relationships as an example, where the way an individual interacts with her relatives, friends, and colleagues may depend on location, time, or means of interaction. Is she at home or at the office? Is it a weekday or weekend? Is she communicating by phone or by Facebook? If all contact events are aggregated into a single network layer, important temporal [19,20] and structural [21] information is inevitably lost. On the other hand, if all different modes of interactions are treated independently in separate network layers, important interplay between the layers is lost. To capture different types of interactions between nodes, researchers have recently introduced multilayer networks together with generalized network methods [22–28], including a generalization of the objective function modularity, to identify groups in multilayer networks [22]. While the generalized null models of modularity are based on Laplacian dynamics [22], they nevertheless favor topological groups with high link density [29], both within and between network layers [30].

III. MODULES IN MULTILAYER NETWORKS: A FLOW APPROACH

To identify modular flows on multilayer networks, we introduce a method based on compression of network flows. The information-theoretic method generalizes the so-called map equation [15] for networks with memory [19] to take advantage of modular flows in multilayer networks. The framework generalizes straightforwardly because the information-theoretic machinery remains the same and only the non-Markovian flow model changes, with memory of the present layer rather than of the previous step. This approach therefore suggests a natural concept of communities in multilayer networks as groups of nodes that capture flows within and across layers for a relatively long time.

We begin by describing how we model the dynamics and then introduce the multiplex map equation. We measure the performance on benchmark networks and contrast with results obtained with the generalization of modularity. Finally, we analyze the modular flow dynamics on two multilayer collaboration networks. Moreover, we have integrated the method in the Infomap software package, which is available online for anyone to use [31].

A. Flow dynamics on multilayer networks

A multilayer network is an efficient representation of a connected system of agents that may interact in different roles, at different times, or by different means. We represent each agent with a physical node, refer to the different means of interaction as different modes, and represent each mode with a network layer. Figure 1 illustrates a multilayer network with four physical nodes and three network layers. We use Latin letters to enumerate the physical nodes, Greek letters to enumerate the network layers, and pairs of Latin and Greek letters to identify node-layer tuples [17]. The node-layer tuples correspond to physical nodes in specific network layers, which we refer to in the following as state nodes [see Figs. 1(c)–1(d)]. Sometimes empirical data allow us to assign weights to both intralayer and interlayer links between state nodes. In such interconnected networks, we have complete information to model dynamics with a random walker that follows links proportional to their weights within and between network layers. Accordingly, movements between state nodes within each layer are Markovian, and movements between physical nodes across layers are non-Markovian.

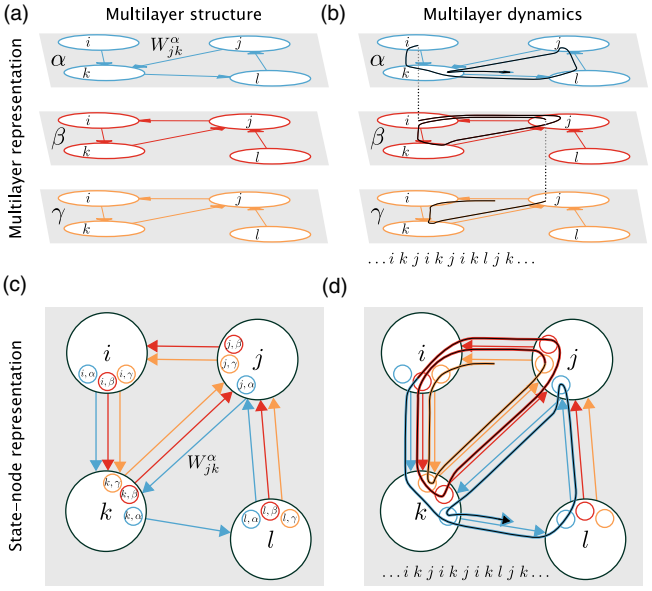

FIG. 1. Modular flow on the benchmark multilayer network. (a) A schematic multilayer network with physical nodes i, j, k, and l, and three layers α, β, and γ in blue, red, and orange, respectively. The physical nodes in each layer are connected with intralayer link weights W^{α}_{jk}. (b) A random walker on the multilayer network moving between the physical nodes in each layer, twice relaxing the layer constraint and following a link from the physical node in any layer. (c) The three layers represented as a multiplex network with physical nodes in black and state nodes i, α in blue, red, and orange. (d) A random walker on the multiplex network moving between the state nodes. While the random walker moves according to the weights between the state nodes, only the physical nodes are considered to be observables, as illustrated by the sequence of physical nodes that the random walker has visited. When the random walker moves along links of the blue layer, it is trapped in the lower-right triangle. When the random walker moves along links of the red or orange layer, it is trapped in the upper-left triangle. As a consequence, the multilayer network has two overlapping modules with respect to flow.

In general, with intralayer adjacency matrix W_{ij}^{β} of layer β and interlayer adjacency matrix $D_i^{\alpha\beta}$ of physical node i, the transition probabilities are

$$\mathcal{P}_{ij}^{\alpha\beta} = \frac{D_i^{\alpha\beta}}{S_i^{\alpha}} \frac{W_{ij}^{\beta}}{s_i^{\beta}}, \tag{1}$$

where $S_i^{\alpha} = \sum_{\beta} D_i^{\alpha\beta}$ are the interlayer out-strengths and $s_i^{\beta} = \sum_j W_{ij}^{\beta}$ are the intralayer out-strengths of node i in layer α and β, respectively [28]. In practice, however, data about interlayer link weights are often scarce. In other words, information about the probability of switching layer is incomplete.

In the absence of empirical interlayer weights, we use random walker dynamics with relax rate r to model movements between layers. In a given step, with probability $1-r$, the random walker moves according to the intralayer links of the state node, and with probability r, the constraint to move in the current layer is relaxed and the random walker moves along any link of the physical node. In this way, the random walker switches from layer α to layer β with probability s_i^{β}/S_i^{α}. These dynamics are described by the transition probabilities

$$\mathcal{P}_{ij}^{\alpha\beta}(r) = (1-r)\delta_{\alpha\beta} \frac{W_{ij}^{\beta}}{s_i^{\beta}} + r \frac{W_{ij}^{\beta}}{S_i}, \tag{2}$$

with $S_i = \sum_{\beta} s_i^{\beta}$ independent of the layer. [It is worth noting that Eq. (2) is equivalent to Eq. (1) when $D_i^{\alpha\beta} = (1-r)\delta_{\alpha\beta} S_i + r s_i^{\beta}$ and $S_i^{\alpha} = \sum_{\beta} s_i^{\beta}$.] A relaxed step on a multilayer network resembles a teleportation step in the PageRank algorithm [8], which allows a random surfer to move freely to a random website and explore the full network. However, a relaxed step only frees the constraints set by the current network layer and allows the random walker to follow a link from node i to node j in any network layer (see Fig. 1). Accordingly, changing the relax rate from 0 to 1 modifies the constraints on the random walker from those that force it to be trapped in disconnected network layers to those that allow it to move more freely on the fully aggregated network. In this way, we can model the important interplay between interconnected network layers. See the Appendix for details about how we model ergodic dynamics.

B. Communities in multilayer networks

There are, in principle, many ways to define communities in multilayer networks [22,32], but the challenge is to construct an effective framework. The challenge may seem daunting since there is still debate about how to define communities in single-layer networks [10], and multilayer networks are inherently more complex with simultaneous and nonlinear coupling between the layers. However, by using the fact that many networks represent constraints on flow in social and biological systems and that multilayer networks are just a more complete description of these constraints, a generalization of flow-based community detection methods follows straightforwardly.

We begin by illustrating how we identify communities in a multilayer network. As an example, we use a social system in which nodes represent individuals and network layers represent interaction processes associated with family, friendship, and work relations, respectively. The constraints on flow in a network layer may give rise to modules with long flow persistence times. Moreover, and importantly, the modules in each network layer may or may not depend on other network layers. For example, if some friends run a business together, their module in the friendship-relations layer will correlate and interplay with their module in the work-relations layer, such that they form a single reinforced module across the two layers. Contrarily, all members of a family may not work together or even interact as friends, such that the family module does not extend across layers. However, if some of the family members run a business together or interact as friends, modules may overlap. In other words, identifying modular flows on multilayer networks captures the fact that individuals can belong to multiple highly interactive communities with limited information transfer between, such that information has long persistence times within communities that may extend into multiple layers. The schematic multilayer network in Fig. 1 illustrates this example. Each layer has a triangle of connected nodes [Fig. 1(a)] that trap flow for a long time [Fig. 1(b)]. The red and the orange network layers interplay more with each other than with the blue network layer. By representing the multilayer network as a multiplex network with state nodes [Fig. 1(c)] and analyzing the dynamics of a random walker on the multiplex network, the community structure with two overlapping modules appears [Fig. 1(d)]: one module across the red and orange layers that captures the interplay between these layers, and a separate module for the blue layer that captures the distinct dynamics in this layer. In general, however, not all modules of a layer need to extend across layers. In the next section, we make this concept of modular flow in multilayer and interconnected networks precise by generalizing the map equation.

C. The multiplex map equation

In short, the map equation takes advantage of the duality in information theory between finding regularities in data and compressing the data [33,34]. It measures the length required to communicate dynamics on a network with a modular description for a given network partition [15]. Therefore, to find the optimal partition, we seek to minimize the description length over all possible network partitions. Accordingly, the network partition that gives the

shortest description length, and compresses the data the most, also best captures the community structure with respect to the dynamics on the network.

We now detail the machinery of the map equation and its natural generalization to multiplex networks. First, here a modular description means that the coding scheme of the map equation grants unique names only to the important structures of the network, the modules. In practice, each entry to a module is assigned a unique code word, while each node visit and module exit is assigned a code word unique only for the particular module. Specifically, one index codebook maps module entries to code words to describe the random walker's movements between modules, and m module codebooks, one for each module, map node visits and module exits to code words to describe the random walker's movements within modules. Importantly, the codebooks are independent, allowing short code words to be reused between them for efficient modular descriptions. Moreover, the code structure with unique code words in modules and of modules, respectively, is equivalent to assuming that the dynamics form an independent and identically distributed process. This equivalence follows from Shannon's source-coding theorem [33], which states that the average description length of an independent and identically distributed random variable X, with events x_i and probability distribution $P(x_i)$, is bounded below by the Shannon entropy $H(X) = -\sum_i P(x_i)\log_2 P(x_i)$. For the modular description of the random walker on the network, the events correspond to node visits and module entries and exits. Accordingly, the average code length of each codebook is derived from the rate of use of each code word. Consequently, to take advantage of the duality between finding regularities and compression, it is not necessary to derive the code words *per se*. Instead, the map equation directly operates on the rates at which a random walker enters and exits modules and visits nodes in the modules, and simply measures the average codelength of each codebook and weights them by how often each one is used.

A simple example illustrates how the machinery works. In the archetype of a modular network, at most weakly connected modules correspond to fully connected cliques of nodes such that a random walker can visit a node with equal probability from any other node in a clique. In other words, node visits are independent and identically distributed. Accordingly, with a module codebook for each clique, and an index codebook for the rare movements between the cliques, the compression is optimal. Any other assignment of nodes to modules would give longer descriptions because the events would no longer be independent and identically distributed or because transitions between modules would be more frequent. In fact, for the clique structure described here, the modular description even achieves the optimal compression over all possible codes, which for random walks on networks is given by the entropy rate of the Markov process obtained by using one codebook for each node with code words for the neighbors [33]. But the constraint of using a modular code structure is necessary for identifying the cliques from optimal compression. And, importantly, while the compression no longer achieves the entropy rate of the Markov process, the machinery nevertheless works for nonclique networks because the partition that best corresponds to a clique structure allows for the best modular compression. Or, from the perspective of the dynamics, maximum compression is achieved when a group of nodes that capture a random walker for a relatively long time is assigned to the same module. In this way, the modular description with unique code words in modules and of modules, respectively, allows for optimal coding of modular dynamics.

Since the map equation expresses the description length in terms of the rates at which a random walker enters and exits modules and visits nodes in the modules, we now derive the rates for a multiplex network. Given a partition M of state nodes i,α assigned to modules $\iota = 1, 2, \ldots, m$, the transition rates at which the random walker enters $q_{\iota\curvearrowleft}$ and exits $q_{\iota\curvearrowright}$ each module take the form

$$q_{\iota\curvearrowleft} = \sum_{\{i,\alpha\}\in J\neq\iota,\{j,\beta\}\in\iota} q_{ij}^{\alpha\beta}, \tag{3}$$

$$q_{\iota\curvearrowright} = \sum_{\{i,\alpha\}\in\iota,\{j,\beta\}\in J\neq\iota} q_{ij}^{\alpha\beta}. \tag{4}$$

However, the original formulation of the map equation was developed for conventional Markovian networks with a single node for each component of the system the network represents. In other words, for the component represented by a node, the node must both define the transition probabilities and constitute the object to be encoded. Much like in a network with memory with its second-order Markov model [19], this constraint is relaxed in a multiplex network. In other words, multiple state nodes of a single physical node capture the transition probabilities of the dynamics, but the coding only captures physical node visits. Therefore, the generalization of the map equation to multiplex networks is simply about separating objects for dynamics and coding. Accordingly, all state nodes of a physical node in a particular module are assigned to a common code word. In this way, the map equation can capture the notion of the multiplex network with its rich and non-Markovian dynamics.

To capture the concept that all state nodes of a physical node in a particular module are assigned to a common code word, the code-word lengths are derived from the rates at which the random walker visits each of the physical nodes in the module. For module codebook ι, the physical node-visit rates are

$$p_{i\in\iota} = \sum_{\{i,\alpha\}\in\iota} p_i^{\alpha}. \tag{5}$$

Moreover, module codebook $\boldsymbol{\iota}$ also has a code word for the exit, derived from the exit rate $q_{\iota\curvearrowright}$ in Eq. (4). We use $p_{\iota\circlearrowleft}$ to denote the sum of these rates and $\mathcal{P}^{\boldsymbol{\iota}} = \{p_{i\in\boldsymbol{\iota}}/p_{\iota\circlearrowleft}\}$ to denote the normalized probability distribution. Similarly, the index codebook has code words for module entries. The code-word lengths are derived from the rates at which the random walker enters each module, $q_{\boldsymbol{\iota}\curvearrowleft}$. We use $q_{\curvearrowleft}$ to denote the sum of these rates and $\mathcal{Q} = \{q_{\boldsymbol{\iota}\curvearrowleft}/q_{\curvearrowleft}\}$ to denote the normalized probability distribution.

We can now express the description length of a random walker in terms of the rates at which it enters and exits modules and visits state nodes of physical nodes in the modules. With the per-step average description length $L(\mathsf{M})$ of the trajectory of an ergodic random walker on a multiplex network, the map equation takes the form

$$L(\mathsf{M}) = q_{\curvearrowleft} H(\mathcal{Q}) + \sum_{\iota=1}^{m} p_{\iota\circlearrowleft} H(\mathcal{P}_{\iota}). \tag{6}$$

This formulation is, on the surface, identical to the standard formulation of the two-level map equation [15], with one important distinction: State nodes of a physical node can be assigned to multiple modules, but if they are assigned to the same module, they are assigned a common code word derived from their total visit rate given by Eq. (5). In colloquial terms in the example above, the distinction corresponds to a scenario in which colleagues who are also friends will refer to each individual by a single name that may be different from what family members use. The seemingly subtle distinction from the standard formulation of the map equation for conventional networks makes, as we show in the following sections, all the difference. The map equation captures the notion of multiplex networks, and multilayer network modules will naturally overlap if the dynamics have such properties.

We have integrated both the two-level and the multilevel multiplex map equation in the Infomap search algorithm available online [31], but here we focus on two-level modular structures, communities.

IV. RESULTS AND DISCUSSION

In this section, we first validate our framework with performance tests on novel multilayer benchmark networks and then analyze two inherently multilayer collaboration networks.

A. Performance tests on multilayer benchmark networks

To test the performance of the information-theoretic and flow-based method, we developed multilayer benchmark networks with modular structure across layers. We followed the standard approach and obtained benchmark networks from a generative model in which nodes are assigned to communities and the probability of drawing a link between two nodes depends on their community assignments [35,36]. While the multiplex map equation can identify modules that independently span across any number of layers, here we consider benchmark networks with community structure in entire layers that either correlate or not. This more easily tractable scenario nevertheless highlights salient features of modular flows. As schematically illustrated in Fig. 1, the scenario corresponds to systems that can be in different modes with dependent network layers. Using the example from above, colleagues would also be friends such that the two layers would have a community structure that is almost the same, yet different from the community structure associated with family relations. Such redundant or complementary information is common in many social and biological networks representing systems that can be in different modes as a whole or slowly change over time [37].

For the mode networks, we first generated T independent Lancichinetti-Fortunato-Radicchi (LFR) benchmark networks [36] for the different modes of the system and then sampled L network layers from each of the mode networks. In the first step, each LFR benchmark network was generated by specifying the degree distribution, the community sizes, and the number of links within and between the communities. Within each community, the links were randomly inserted according to the configuration model [38], and the same model was used to insert links between the communities. Specifically, we used LFR benchmark networks with 128 nodes and 4 communities, each with 32 nodes with average degree 16, and the fraction of intercommunity links set to 0.05. In the second step, to sample a link of the mode network once, on average, we sampled the network layers by including each link with probability $1/L$. Each multilayer benchmark network thus comprises $T \times L$ layers, with T sets of L dependent layers.

Figure 2 schematically illustrates a multilayer benchmark network with $T = L = 2$. The challenge is to reveal the community structure of each mode network, simultaneously revealing the community structure in each layer and identifying the mode network from which the layer was sampled. To make the test realistic, we only provided the algorithm with the $T \times L$ network layers and did not input any information about the number of mode networks T or about how or in which order the layers were sampled. In the small example illustrated in Fig. 2, generalized modularity [22] correctly identifies the communities in each layer but fails to identify the communities in the two original mode networks. Contrarily, the multiplex map equation, here with relax rate $r = 0.15$, identifies both the communities in each layer and the communities in the mode networks.

The multiplex map equation can accurately identify multilayer communities. To test the performance more

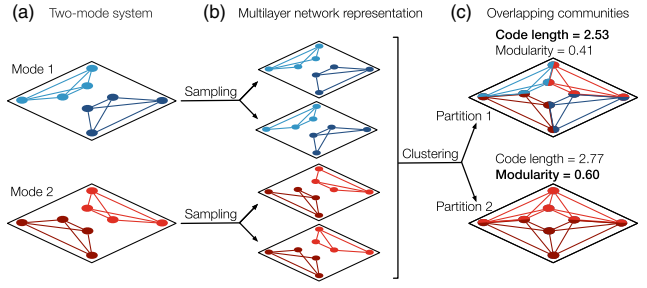

FIG. 2. Overlapping communities in multilayer benchmark networks. We generate the multilayer networks in two steps. (a) First, we generate T LFR benchmark networks with well-defined communities, here illustrated with two network modes in blue and red. (b) Then, we sample L network layers from each mode network, here illustrated with four network layers in total. (c) Each state node in the multilayer network is classified in a community, such that communities of physical nodes may overlap. In partition 1, each state node is correctly classified. In partition 2, however, the light and dark color shades are assigned to the same module, respectively. While these communities provide the correct partition of each slice, they fail to capture the communities of the two original mode networks. Generalized modularity favors partition 2, whereas the multiplex map equation favors partition 1.

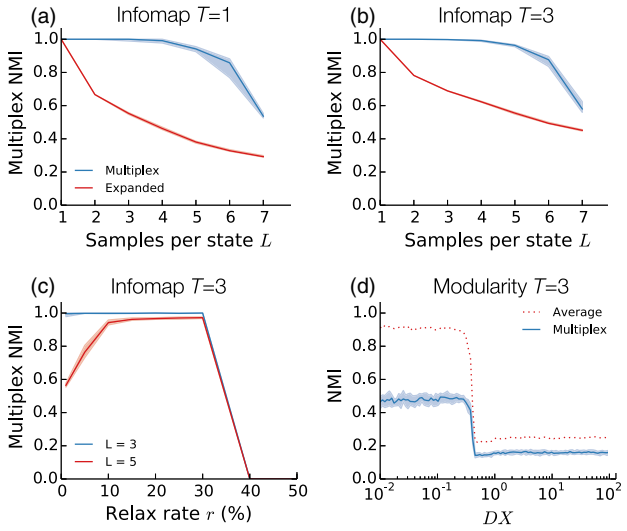

FIG. 3. Performance test on multilayer benchmark networks. (a,b) Performance of multiplex Infomap (Multiplex) compared with Infomap applied to the expanded network with state nodes interpreted as physical nodes (Expanded) and to each network layer separately (Single) as a function of number of network layers for 1 and 3 mode networks, respectively. We used relax rate $r = 0.15$ and quantified the performance by the NMI between the planted and obtained partitions of state nodes. (c) Performance of multiplex Infomap as a function of the relax rate r. (d) Performance of generalized modularity optimization for $T = L = 3$ as a function of the interlayer coupling DX, measured both as the NMI of state nodes (Multiplex) and averaged across network layers (Average).

systematically, we generated multilayer benchmark networks with different numbers of mode networks T (between 1 and 3) and network layers per mode network L (between 1 and 7), and we applied the normalized mutual information (NMI) [39,40] to state nodes (multiplex NMI). In this way, we can quantify how well the method captures the multilayer communities. Figures 3(a) and 3(b) show the results for relax rate $r = 0.15$. Optimization of the multiplex map equation with Infomap, multiplex Infomap for short, accurately identifies the communities of the original mode networks for up to 5–6 network layers per mode network. Contrarily, the standard Infomap applied on each layer separately or on the supra-adjacency representation of the multilayer network with all state nodes interpreted as physical nodes [26] only succeeded for one layer per mode network. Similarly, generalized modularity [22] does not identify this type of planted community across network layers [see Fig. 3(d)] because it uses a null model only for intralayer links and merely a coupling parameter between layers [22,30]. As a result, merging different communities across layers always improves the modularity score. For example, by measuring the performance on each layer separately, by simply averaging the multiplex NMI applied to each layer separately (average NMI), Fig. 3(d) shows that generalized modularity, besides optimization and resolution-limit problems, accurately captures the communities at each layer. Accordingly, as illustrated in Fig. 2(c), generalized modularity cannot separate dependent communities from independent communities across layers. In Ref. [45], Fig. S1 shows the corresponding analysis for Infomap. Also in this test, the multiplex Infomap gives better results than Infomap applied on each layer separately because Infomap tends to overestimate the number of clusters in the sparse network layers. In other words, multiplex Infomap can use information across layers for better performance on individual layers. In any case, only by acknowledging the multiplex nature of the benchmark networks is it possible to accurately identify their multilayer communities.

The results for multiplex Infomap are only weakly dependent on the relax rate [see Fig. 3(c)], although the exact range depends on the relative constraints on flow manifested in network layers of the same and different mode networks (see Figs. S1 and S2 in Ref. [45]). When nothing else is stated, we use $r = 0.15$ throughout our analysis. With this relax rate, the random walker stays in the same network layer for about six steps.

Overall, we were not able to recover multilayer communities by treating the multilayer network as one large network, as multiple disconnected networks, or as multiple networks connected with a coupling parameter without a proper null model. We conclude that the key discriminating factor is the map equation's ability to capture the important notion of multiplex networks that sets of state nodes across layers represent the very same physical objects.

B. Multilayer community structure of collaboration networks

We analyzed two inherently multilayer collaboration networks, the Pierre Auger Collaboration of physicists and a sample from the arXiv of researchers working on networks. The Pierre Auger Collaboration is a group of hundreds of theoretical and experimental scientists worldwide working at the Pierre Auger Observatory, the largest observatory of ultra-high-energy cosmic rays [41]. The collaborators work together in different research topics on specific tasks, e.g., source detection, mass composition, experimental enhancements, and shower reconstruction, etc. Scientists within the Collaboration may work on one or more tasks, and every year, hundreds of internal technical reports are submitted to the repository. [46] With access to author lists and keywords, we reconstructed the inherently multilayer collaboration network in which nodes represent scientists, links indicate collaboration between scientists, and layers represent tasks (see Table I). We considered all submissions between 2010 and 2012 and assigned each report to $L = 16$ layers according to its keywords and its content, with manual disambiguation to avoid spurious results from an automated process. For each report with more than one author, for each layer in which the report was classified, and between any pair of the $N = 514$ co-authors, we added a weight $1/(L(N-1))$ to the weighted, undirected, and multilayer network. In this way, the sum of all link weights of an author across all layers is simply the number of reports written by the author. We built the arXiv [47] multilayer network in the same way, but instead of tasks, we used arXiv categories for layers (see Table II). To restrict the analysis to a well-defined topic of research, we only included papers with "networks" in the title or abstract up to May 2014. We found 12,019 articles from 14,488 authors, whose names have been disambiguated by using heuristics. Because some categories or tasks are more related than others, communities naturally emerge across layers when groups of scientists work on interdisciplinary projects or several tasks simultaneously.

TABLE I. The Pierre Auger Observatory: Each task defines a layer in the multilayer co-authorship network.

Layer	Task
1	Neutrinos
2	Detector
3	Enhancements
4	Anisotropy
5	Point source
6	Mass composition
7	Horizontal
8	Hybrid reconstruction
9	Spectrum
10	Photons
11	Atmospheric
12	SD reconstruction
13	Hadronic interactions
14	Exotics
15	Magnetic
16	Astrophysical scenarios

TABLE II. The arXiv repository: Each category defines a layer in the multilayer co-authorship network.

Layer	Category
1	physics.soc-ph
2	physics.data-an
3	physics.bio-ph
4	math-ph
5	math.OC
6	cond-mat.dis-nn
7	cond-mat.stat-mech
8	q-bio.MN
9	q-bio
10	q-bio.BM
11	nlin.AO
12	cs.SI
13	cs.CV

The collaboration networks show a highly overlapping modular organization. In Fig. 4(a), we show the largest connected component of the Pierre Auger Collaboration network, including more than 90% of the scientists, and their assignments into highly overlapping modules. Truly multilayer nodes, i.e., those corresponding to scientists active in more than one task, dominate the core of the network in this visualization, whereas single-task scientists are more peripheral nodes. For example, the multilayer analysis reveals strong groups of collaboration across the tasks of "point source," "anisotropy," and "magnetic," [see Fig. 5(a)]. Figure 4(b) shows that essential information about the overlapping modular organization is washed out when dynamics are modeled with $r = 1.0$ or in the aggregated network (not shown in the figure because, qualitatively, it provided the same results), and scientists are assigned to a few overlapping communities ($r = 1.0$) or one community only (aggregated network). Without mentioning names, we find scientists who are indisputably active in several different tasks, with variegated collaboration patterns captured only when dynamics are modeled with $r = 0.15$, whereas for $r = 1.0$, the scientists are grouped in single nonoverlapping communities. In another case, we find two colleagues who work at nearby institutions within the same city and with highly overlapping interests and collaborations. For $r = 0.15$, they are assigned to highly overlapping modules across tasks, whereas for $r = 1.0$, they are assigned to different nonoverlapping partitions. Only by maintaining the multilayer structure were we able to reveal the actual collaboration structure. Similarly, Fig. 5(b) shows that communities also extend across layers in the arXiv collaboration network.

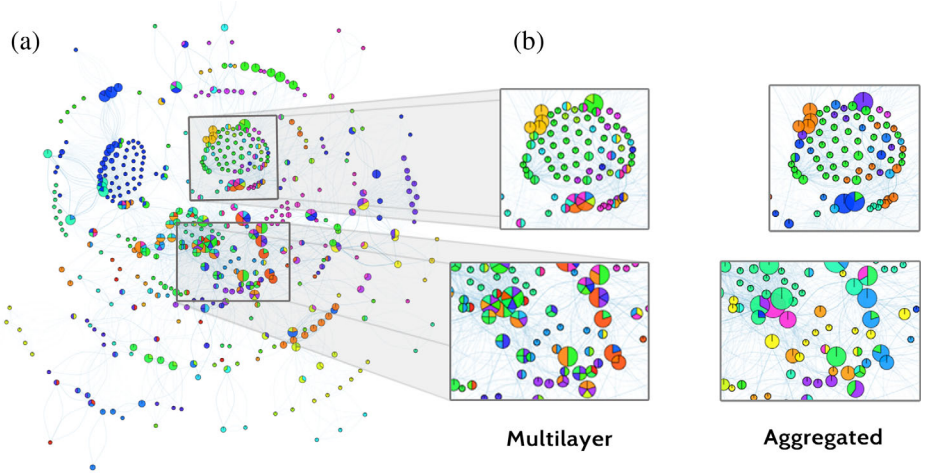

FIG. 4. Community structure in the Pierre Auger Collaboration network. (a) The overlapping community structure revealed by the multiplex map equation with relax rate $r = 0.15$. Nodes for scientists are colored according to their module assignments, with node sizes proportional to the number of tasks in which they were active. Specifically, the area of a colored pie-chart slice is proportional to the number of tasks in which the corresponding scientist is active. (b) Subsets of nodes with direct comparison with the overlapping community structure obtained from dynamics with $r = 1.0$.

Whereas communities typically only extend a few layers in the Pierre Auger Collaboration network, communities in the arXiv network can extend over multiple layers. This means that scientists are rather task specific in the Pierre Auger Collaboration, whereas researchers working on networks often are involved in interdisciplinary projects, although computer vision and mathematics seem to be less interdisciplinary topics. In any case, the multilayer networks analyzed with the map equation capture the fact that scientists can simultaneously work in different groups on different topics.

Table III summarizes the multilayer effects of community detection with the map equation framework. For easy comparison, we contrast multilayer results obtained with dynamics modeled with relax rate $r = 0.15$ with results obtained with relax rate $r = 1.0$. The latter maximum relax rate corresponds to completely washed-out multilayer information, but, unlike the aggregated networks, it allows multiplex Infomap to assign nodes to multiple modules. For both the Pierre Auger and arXiv networks, we find that flow is confined in smaller and more overlapping modules. For a physics explanation, we also measure this effect in terms of the persistence gain in modules. For modules obtained with $r = 0.15$, the persistence gain quantifies how much longer a random walker stays within the modules when dynamics are modeled with $r = 0.15$ compared with $r = 1.0$. When a random walker only moves freely between layers in one step out of about six, compared with free movements between layers in any step, we find that its chance to stay within the same module increases by 25% and 13% in the

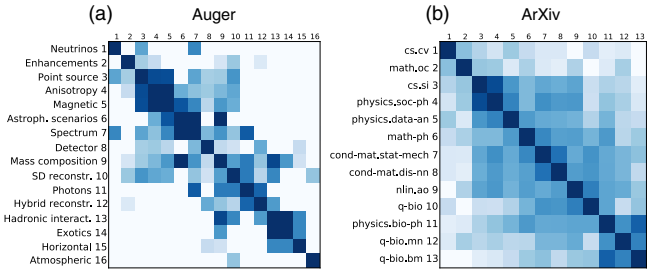

FIG. 5. Real multilayer networks with communities across network layers. The heat maps show the similarities between network layers, measured as the fraction of state nodes in different network layers that are assigned to the same communities.

TABLE III. Summary of multilayer effects on community detection.

	Synthetic networks		Real networks	
	$T = 1$	$T = 3$	Auger	arXiv
Number of nodes n	256	256	514	14,488
Number of links l	1,400	4,000	12,964	70,350
Number of layers L_{tot}	3	9	16	13
NMI, r_{15} vs r_{100}	1.0	0.0	0.74	0.92
Effective module size, r_{15}	32	11	10	13
Effective module size, r_{100}	32	128	16	17
Module assignment, r_{15}	1.0	3.0	1.4	1.2
Module assignment, r_{100}	1.0	1.0	1.1	1.0
Persistence gain (%)	0	163	25	13
Compression gain (%)	0	32	26	22

Synthetic networks with $L = 3$ layers per state T. Modeled dynamics denoted r_{15} and r_{100} for relax rates 0.15 and 1.0, respectively. Effective module size measured as $n/2^{H(\mathcal{S})}$, where $H(\mathcal{S})$ is the entropy of the distribution of module sizes in terms of their flow volumes. Persistence and compression gains for dynamics modeled with r_{15} compared with r_{100}, and with modular solution obtained for r_{15}. All figures are significant.

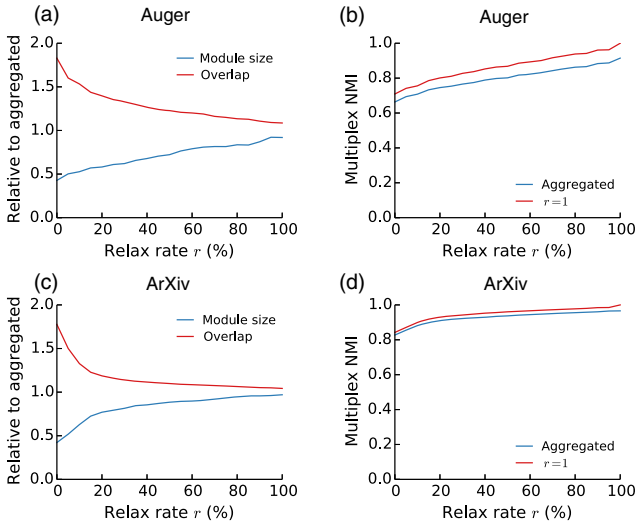

FIG. 6. Aggregation is responsible for significant changes in the community structure. The panels show module sizes, number of assignments per node (overlap), and NMI for communities revealed from the multilayer and the aggregated networks in the Pierre Auger Collaboration (top panels) and the arXiv (bottom panels) networks. For a given relax rate, the NMI measures the similarity between the obtained partition and the partitions obtained from the aggregated topology (blue curve) and the aggregated dynamics at relax rate $r = 1.0$ (red curve), respectively.

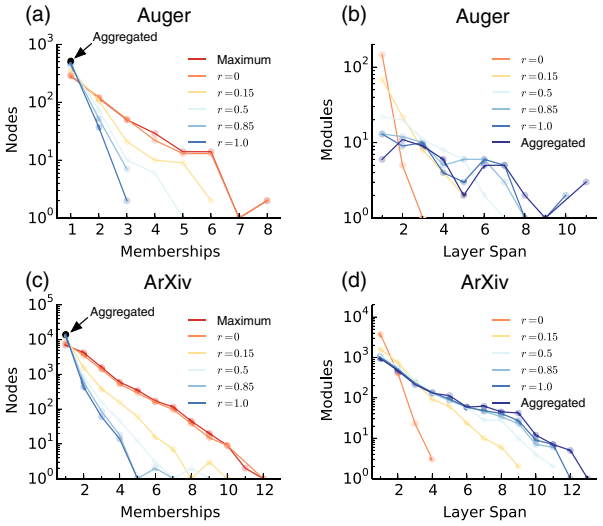

FIG. 7. Modules span more layers and overlap less with increased relax rate. Panels (a) and (c) show the number of nodes as a function of the memberships they have, i.e., the number of modules the nodes belong to, for several values of the relax rate. For each node, the *maximum* number of memberships is the number of layers where the node appears: This is very close to what we get for $r = 0$, although the curves are not exactly the same. For $r = 1$, we still have some overlap compared to the aggregated one, which is just a single point because each node belongs to one single module. Panels (b) and (d) show the number of modules that are present in the number of layers indicated on the x axis. For small values of r, the modules tend to be localized in fewer layers.

Pierre Auger and arXiv networks, respectively. As a result of this persistence gain, the modular description of a random walker's trajectory can be significantly compressed in both networks.

We also investigate how the relax rate allow us to change the resolution of the module and the overlap across layers. Figures 6 and 7 show the differences between the partitions found with the multilayer and the aggregated approach. At an increasing relax rate, the random walker becomes less and less localized in a specific layer. Accordingly, the NMI between the multilayer and the aggregated solutions increases. For $r = 1$, the walker moves freely between layers, but the NMI does not equal 1 because the multilayer solution still allows for overlap (the optimization algorithm we used on the aggregated solution does not identify overlaps by construction).

In both data sets, with increasing relax rate, we find bigger modules (module size increases) which span more layers, and fewer community assignments per physical node (overlap decreases). The module size is defined as the number of nodes divided by the effective number of modules. The effective number of modules is defined as 2^H, where H is the Shannon entropy of the partition.

Since compressing data is dual to finding regularities in the data [33,42], the multiplex map equation applied to the multilayer representation allows us to discover patterns that are absent in the aggregated network. Evidently, these patterns contain essential information about the constraints on flow through the systems.

V. CONCLUSIONS

In summary, compared with conventional network analysis, multiplex Infomap applied to the studied multilayer networks uncovers interplay between network layers and reveals smaller modules with more overlap that better capture the actual organization. Shoehorning multiplex networks into conventional community-detection algorithms can obscure important structural information, and earlier attempts at generalizing conventional community-detection methods to identify communities across layers have proven problematic. In contrast, thanks to the map equation's intrinsic ability to capture that sets of nodes across layers represent the very same physical objects in multiplex networks, the framework generalizes straightforwardly. In the absence of empirical interlayer links, here we have modeled the dynamics between layers. However, interlayer interaction data would provide further important information about the organization of social and biological systems, thus calling for more empirical work.

ACKNOWLEDGMENTS

A. A. and M. D. D. were supported by the European Commission FET-Proactive project PLEXMATH (Grant No. 317614) and the Generalitat de Catalunya

2009-SGR-838. A. A. also acknowledges partial financial support from the European Commission FET-Proactive project MULTIPLEX (Grant No. 317532), ICREA Academia, and the James S. McDonnell Foundation. M. R. was supported by the Swedish Research Council Grant No. 2012-3729. The authors acknowledge all members of the Pierre Auger Collaboration for kindly providing access to the metadata of its repository for internal technical reports, Dr. M. Settimo for kindly helping to classify all reports to the proper task(s), and P. J. Mucha, M. A. Porter, M. Bazzi, and L. Jeub for fruitful discussions.

APPENDIX: DYNAMICS ON MULTILAYER NETWORKS

The rationale behind the multiplex map equation is simple: Encode the trajectory between physical nodes of a random walker that navigates between state nodes in different layers [see Fig. 1(c)]. For a modular description, the trajectory is encoded with unique code words on all modules and all physical nodes within each module, respectively. We are only interested in the codelengths and can derive them from the stationary distribution of the random walker. The stationary distribution on the state nodes can be derived from the transition probabilities $\mathcal{P}_{ij}^{\alpha\beta}$ described in Eq. (1) for interconnected networks with empirical interlayer link weights and in Eq. (2) for multilayer networks with interlayer link weights modeled with relaxation parameter r. Assuming that the stationary distribution of state node i,α is p_i^α, it can, in principle, be derived from the recursive system of equations

$$p_i^\alpha = \sum_{j,\beta} p_j^\beta \mathcal{P}_{ji}^{\beta\alpha}. \tag{A1}$$

However, to guarantee a unique ergodic solution in directed networks, we use teleportation at a low rate τ to state nodes proportional to their intralayer in-strength [43]. To reduce the smoothening effect of teleportation and make the results more robust to the teleportation parameter τ, we use unrecorded teleportation steps and recorded steps along links [43]. We obtain the recorded visit rates by first calculating the stationary distribution with teleportation to state nodes proportional to their out-strength,

$$\tilde{p}_i^\alpha = (1-\tau)\sum_{j,\beta} p_j^\beta \mathcal{P}_{ji}^{\beta\alpha} + \tau \frac{s_i^\alpha}{\sum_{i,\alpha} s_i^\alpha}, \tag{A2}$$

with the power-iteration method [44]. Then, we derive the recorded steps along links $q_{ji}^{\beta\alpha}$ and nodes p_i^α in a subsequent step,

$$q_{ji}^{\beta\alpha} = \tilde{p}_j^\beta \mathcal{P}_{ji}^{\beta\alpha}, \tag{A3}$$

$$p_i^\alpha = \sum_{j,\beta} q_{ji}^{\beta\alpha}. \tag{A4}$$

We use teleportation rate $\tau = 0.15$ throughout our analysis of directed networks, but the results are robust to variation of τ in a wide range around this value. For undirected networks, results are independent of τ.

[1] J. Scott, *Social Network Analysis*, Sociology **22**, 109 (1988).
[2] H. Kitano, *Systems Biology: A Brief Overview*, Science **295**, 1662 (2002).
[3] M. E. J. Newman, *The Structure and Function of Complex Networks*, SIAM Rev. **45**, 167 (2003).
[4] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, *The Architecture of Complex Weighted Networks*, Proc. Natl. Acad. Sci. U.S.A. **101**, 3747 (2004).
[5] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang, *Complex Networks: Structure and Dynamics*, Phys. Rep. **424**, 175 (2006).
[6] A. L. Barabási and R. Albert, *Emergence of Scaling in Random Networks*, Science **286**, 509 (1999).
[7] Mark E. J. Newman, *Assortative Mixing in Networks*, Phys. Rev. Lett. **89**, 208701 (2002).
[8] S. Brin and L. Page, *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Comput. Networks ISDN Syst. **30**, 107 (1998).
[9] A. Vespignani, *Modelling Dynamical Processes in Complex Socio-technical Systems*, Nat. Phys. **8**, 32 (2012).
[10] S. Fortunato, *Community Detection in Graphs*, Phys. Rep. **486**, 75 (2010).
[11] M. E. J. Newman, *Modularity and Community Structure in Networks*, Proc. Natl. Acad. Sci. U.S.A. **103**, 8577 (2006).
[12] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, *Finding Statistically Significant Communities in Networks*, PLoS One **6**, e18961 (2011).
[13] T. P. Peixoto, *Parsimonious Module Inference in Large Networks*, Phys. Rev. Lett. **110**, 148701 (2013).
[14] I. Simonsen, K. Astrup Eriksen, S. Maslov, and K. Sneppen, *Diffusion on Complex Networks: A Way to Probe Their Large-Scale Topological Structures*, Physica (Amsterdam) **336A**, 163 (2004).
[15] M. Rosvall and C. T. Bergstrom, *Maps of Random Walks on Complex Networks Reveal Community Structure*, Proc. Natl. Acad. Sci. U.S.A. **105**, 1118 (2008).
[16] J. C. Delvenne, S. N. Yaliraki, and M. Barahona, *Stability of Graph Communities Across Time Scales*, Proc. Natl. Acad. Sci. U.S.A. **107**, 12755 (2010).
[17] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, *Multilayer Networks*, J. Complex Networks **2**, 203 (2014).
[18] S. Boccaletti, G. Bianconi, R. Criado, C. I. Del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin, *The Structure and Dynamics of Multilayer Networks*, Phys. Rep. **544**, 1 (2014).
[19] M. Rosvall, A. V. Esquivel, A. Lancichinetti, J. D. West, and R. Lambiotte, *Memory in Network Flows and Its Effects on*

Spreading Dynamics and Community Detection, Nat. Commun. **5**, 4630 (2014).
[20] P. Holme and J. Saramäki, *Temporal Networks*, Phys. Rep. **519**, 97 (2012).
[21] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin, *Catastrophic Cascade of Failures in Interdependent Networks*, Nature (London) **464**, 1025 (2010).
[22] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, *Community Structure in Time-Dependent, Multiscale, and Multiplex Networks*, Science **328**, 876 (2010).
[23] V. Nicosia, G. Bianconi, V. Latora, and M. Barthelemy, *Growing Multiplex Networks*, Phys. Rev. Lett. **111**, 058701 (2013).
[24] F. Radicchi and A. Arenas, *Abrupt Transition in the Structural Formation of Interconnected Networks*, Nat. Phys. **9**, 717 (2013).
[25] A. Cardillo, J. Gómez-Gardeñes, M. Zanin, M. Romance, D. Papo, F. del Pozo, and S. Boccaletti, *Emergence of Network Features from Multiplexity*, Sci. Rep. **3**, 1344 (2013).
[26] S. Gómez, A. Díaz-Guilera, J. Gómez-Gardeñes, C. J. Pérez-Vicente, Y. Moreno, and A. Arenas, *Diffusion Dynamics on Multiplex Networks*, Phys. Rev. Lett. **110**, 028701 (2013).
[27] M. De Domenico, A. Solè-Ribalta, E. Cozzo, M. Kivelä, Y. Moreno, M. A Porter, S. Gòmez, and A. Arenas, *Mathematical Formulation of Multi-layer Networks*, Phys. Rev. X **3**, 041022 (2013).
[28] M. De Domenico, A. Solé-Ribalta, S. Gómez, and A. Arenas, *Navigability of Interconnected Networks under Random Failures*, Proc. Natl. Acad. Sci. U.S.A. **111**, 8351 (2014).
[29] R. Lambiotte, J.-C. Delvenne, and M. Barahona, *Laplacian Dynamics and Multiscale Modular Structure in Networks*, arXiv:0812.1770.
[30] G. Petri and P. Expert, *Temporal Stability of Network Partitions*, Phys. Rev. E **90**, 022813 (2014).
[31] D. Edler and M. Rosvall, *The Infomap software package*, http://www.mapequation.org.
[32] C. W. Loe and H. J. Jensen, *Comparison of Communities Detection Algorithms for Multiplex*, arXiv:1406.2205.
[33] C. E. Shannon, *A Mathematical Theory of Communication*, Bell Syst. Tech. J. **27**, 379 (1948).
[34] P. D. Grünwald, *The Minimum Description Length Principle* (MIT Press, Cambridge, Massachusetts, 2007).
[35] M. Girvan and M. E. J. Newman, *Community Structure in Social and Biological Networks*, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).
[36] A. Lancichinetti, S. Fortunato, and F. Radicchi, *Benchmark Graphs for Testing Community Detection Algorithms*, Phys. Rev. E **78**, 046110 (2008).
[37] M. De Domenico, V. Nicosia, A Arenas, and V. Latora, *Structural Reducibility of Multilayer Networks*, arXiv: 1405.0425 [Nat. Commun. (to be published)].
[38] M. Molloy and B. Reed, *A Critical Point for Random Graphs with a Given Degree Sequence*, Random Struct. Algor. **6**, 161 (1995).
[39] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, *Comparing Community Structure Identification*, J. Stat. Mech. 2005 P09008.
[40] M. Meilă, *Comparing Clusterings: An Information Based Distance*, J. Multivariate Anal. **98**, 873 (2007).
[41] Pierre Auger Collaboration, in Contributions of the Pierre Auger Collaboration to the 33rd International Cosmic Ray Conference, Rio de Janeiro, Brazil, July 2013.
[42] J. Rissanen, *Modeling by Shortest Data Description*, Automatica **14**, 465 (1978).
[43] R. Lambiotte and M. Rosvall, *Ranking and Clustering of Nodes in Networks with Smart Teleportation*, Phys. Rev. E **85**, 056107 (2012).
[44] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Vol. 3 (JHU Press, Baltimore, Maryland, 2012).
[45] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevX.5.011027 for additional numerical experiments with toy models and high-resolution images of empirical community structures.
[46] Official web page of the Pierre Auger Observatory, http://www.auger.org. Note that the repository is not publicly accessible and privacy policies have been considered, anonymizing the data by assigning a random numerical integer to each author.
[47] Official web page of the arXiv repository, http://www.arxiv.org.