Multilayer Stochastic Block Models Reveal the Multilayer Structure of Complex Networks

Toni Vallès-Català,[1] Francesco A. Massucci,[1] Roger Guimerà,[2,1,*] and Marta Sales-Pardo[1,†]
[1]*Departament d'Enginyeria Química, Universitat Rovira i Virgili, 43007 Tarragona, Catalonia, Spain*
[2]*Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Catalonia, Spain*
(Received 30 June 2015; revised manuscript received 23 December 2015; published 31 March 2016)

In complex systems, the network of interactions we observe between systems components is the aggregate of the interactions that occur through different mechanisms or layers. Recent studies reveal that the existence of multiple interaction layers can have a dramatic impact in the dynamical processes occurring on these systems. However, these studies assume that the interactions between systems components in each one of the layers are known, while typically for real-world systems we do not have that information. Here, we address the issue of uncovering the different interaction layers from aggregate data by introducing multilayer stochastic block models (SBMs), a generalization of single-layer SBMs that considers different mechanisms of layer aggregation. First, we find the complete probabilistic solution to the problem of finding the optimal multilayer SBM for a given aggregate-observed network. Because this solution is computationally intractable, we propose an approximation that enables us to verify that multilayer SBMs are more predictive of network structure in real-world complex systems.

DOI: 10.1103/PhysRevX.6.011036 Subject Areas: Complex Systems, Statistical Physics

I. INTRODUCTION

The development of tools for the analysis of real-world complex networks has significantly advanced our understanding of complex systems in fields as diverse as molecular and cell biology [1], neuroscience [2], biomedicine [3,4], ecology [5,6], economics [7], and sociology [8]. One of the main successes of the network approach has been to unravel the relationship between the modular organization of interactions within a complex system [9] and the function and temporal evolution of the system [10–13]. As a result, a large body of research has been devoted to the detection of the modular structure (or community structure) of complex networks, that is, to the division (partition) of the nodes of the network into densely connected subgroups [14].

Stochastic block models (SBMs) [15–17] are a class of probabilistic generative network models that provide a more general description of the (mesoscopic) group structure of real-world networks than modular models. In SBMs, nodes are assumed to belong to groups and connect to each other with probabilities that depend only on their group memberships. The simple mathematical form of SBMs has enabled not only the identification of generalized community structures in networks [17–26], but also has made network inference a predictive tool to detect missing and spurious links in empirical network data [27], to predict human decisions [28,29] and the appearance of conflict in work teams [30], or for the identification of unknown interactions between drugs [31].

While these approaches have pushed forward our understanding of complex network structure, a limitation is that they rely on the premise that there is a single mechanism that describes the connectivity of the network, even though we know that real-world networks are often the result of processes occurring on different "layers" (for example, social networks encompass relationships that arise on the familiar layer and relationships that arise in the professional layer) [32]. Moreover, it is increasingly clear that the multilayer structure of complex networks can have a dramatic impact on the dynamical processes that take place on them [33–38]. Unfortunately, we often lack information about the different layers of interaction and can only observe projections of these multilayer interactions into an aggregate network in which all links are equivalent.

Here, we precisely address the problem of unraveling the underlying multilayer structure in real-world networks. First, we introduce the family of multilayer SBMs that generalizes single-layer SBMs to situations where links arise in different layers and are aggregated. Although there have been proposals to extend the concept of modularity to multilayer networks [39], ours represents a pioneering attempt to extend generative group-based models to multilayer systems, and to study those models rigorously using tools from statistical physics. Our approach is also different from so-called latent feature models [40–42] in that SBMs allow us to answer the fundamental question of whether an

*Corresponding author.
roger.guimera@urv.cat
†Corresponding author.
marta.sales@urv.cat

observed network is the outcome of a multilayer process, while in latent feature models it is impossible to disentangle the contributions of each layer.

Second, we give the probabilistically complete solution to the problem of inferring the optimal multilayer SBM for a given aggregate network. Because this solution is computationally intractable, we propose an approximation that enables us to objectively address the question of whether an observed network is likely to be the projection of multiple layers. The analysis of complex networks from different contexts suggests that many real-world networks are indeed projections.

II. MULTILAYER STOCHASTIC BLOCK MODELS

In our approach, nodes interact in different layers. In each one of these layers, $\ell = 1, \ldots, L$, we define a SBM as follows: each node i belongs to a specific group σ_i^ℓ, and links between pairs of nodes belonging to groups α and β, respectively, in layer ℓ exist with probability $q_{\alpha\beta}^\ell$. The observed adjacency matrix $A^{\mathcal{O}}$ is an aggregate that results from the combination of the links in each of the layers, and where all information of the layers has been lost (Fig. 1). We call this model the multilayer SBM.

Here, we consider the simplest multilayer case and set $L = 2$. In such case, there are two combinations with a plausible physical interpretation: (i) the *AND* combination of layers, in which $A_{ij}^{\mathcal{O}} = 1$ if, and only if, nodes i and j are connected in both layers [Fig. 1(a)], and (ii) the *OR* combination of layers, in which $A_{ij}^{\mathcal{O}} = 1$ if i and j are connected in at least one layer [Fig. 1(b)]. For example, the AND model is a plausible model for *in vivo* protein interactions, because in order for proteins to interact in the cell it is necessary for them to be capable of physically interacting (that is, to be linked in the layer of *in vitro* physical interactions) and to be expressed simultaneously in the same cellular compartment (that is, to be linked in the coexpression layer). The OR model is a plausible model for the effective online social network through which *memes* spread [43], because some people use Facebook to share memes, others use Twitter, and others use both.

In principle, we would like to identify which is the pair of partitions $(\mathcal{P}_1, \mathcal{P}_2)$ (in layers 1 and 2, respectively) that best describe the observed aggregate topology, which has no information about the layers. The probabilistically complete way to solve this problem is to obtain the joint probability $P(\mathcal{P}_1, \mathcal{P}_2|A^{\mathcal{O}})$ that $\mathcal{P}_1$ and $\mathcal{P}_2$ are the true partitions of the nodes given the aggregate observed network. This distribution is given by

$$P(\mathcal{P}_1, \mathcal{P}_2|A^{\mathcal{O}}) \propto \int DQ_1 \int DQ_2 P(A^{\mathcal{O}}|Q_1, Q_2, \mathcal{P}_1, \mathcal{P}_2) \times P(Q_1, Q_2, \mathcal{P}_1, \mathcal{P}_2), \quad (1)$$

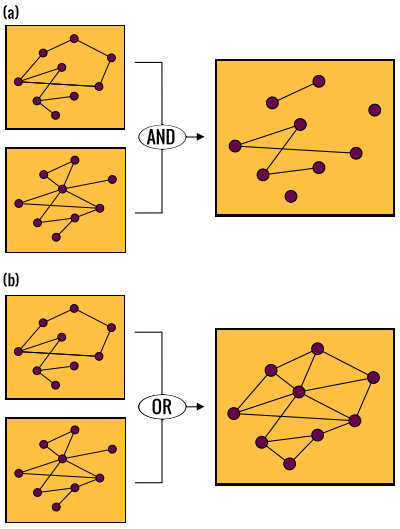

FIG. 1. Network aggregation mechanisms. In aggregated multilayer networks, different networks containing the same nodes but with different adjacency matrices are combined into an observed network with adjacency matrix $A^{\mathcal{O}}$, where all information about the original layers has been lost. We consider two aggregation mechanisms of two-layer networks with adjacency matrices A^1 and A^2: (a) AND aggregation, in which $A_{ij}^{\mathcal{O}} = A_{ij}^1 A_{ij}^2$, so that $A^{\mathcal{O}} = 1$ if, and only if, i and j are connected in both layers, and (b) OR aggregation in which $A_{ij}^{\mathcal{O}} = 1 - (1 - A_{ij}^1)(1 - A_{ij}^2)$, so that $A^{\mathcal{O}} = 1$ if i and j are connected in at least one layer.

where Q_ℓ is a matrix whose elements $q_{\alpha\beta}^\ell$ represent the probability that a link exists between a pair of nodes belonging to groups α and β in layer ℓ, and $\int DQ_\ell \equiv \prod_{\alpha \le \beta} \int_0^1 dq_{\alpha\beta}^\ell$ is the integral over all possible values of these probabilities.

This integral can be computed both for AND combinations and for OR combinations of the two layers; for clarity, we show the calculation for the AND model and discuss the OR model in Appendix B. Because in a SBM each link is independent of each other and in the AND model a link has to be present in both layers to appear in the observed aggregate network $A^{\mathcal{O}}$, the likelihood for an AND model is

$$P_{\text{AND}}(A^{\mathcal{O}}|Q_1, Q_2, \mathcal{P}_1, \mathcal{P}_2) = \prod_{\substack{[\alpha \le \beta] \\ [\gamma \le \delta]}} (q_{\alpha\beta}^1 q_{\gamma\delta}^2)^{n_{\alpha\beta\gamma\delta}^1} (1 - q_{\alpha\beta}^1 q_{\gamma\delta}^2)^{n_{\alpha\beta\gamma\delta}^0}, \quad (2)$$

where $n_{\alpha\beta\gamma\delta}^1$ is the number of links between pairs of nodes that are in groups α and β, respectively, in layer 1, and in groups γ and δ, respectively, in layer 2 ($n_{\alpha\beta\gamma\delta}^1 = \sum_{i<j} A_{ij}^{\mathcal{O}} \delta_{\sigma_i^1 \alpha} \delta_{\sigma_j^1 \beta} \delta_{\sigma_i^2 \gamma} \delta_{\sigma_j^2 \delta}$), and $n_{\alpha\beta\gamma\delta}^0$ is the number of no links between such pairs of nodes [$n_{\alpha\beta\gamma\delta}^0 = \sum_{i<j} (1 - A_{ij}^{\mathcal{O}}) \delta_{\sigma_i^1 \alpha} \delta_{\sigma_j^1 \beta} \delta_{\sigma_i^2 \gamma} \delta_{\sigma_j^2 \delta}$].

Assuming a uniform distribution for the prior $P(Q_1, Q_2, \mathcal{P}_1, \mathcal{P}_2) = \text{const}$ [27,44], we can plug Eq. (2) into Eq. (1) and integrate to find (Appendix A)

$$P_{\text{AND}}(\mathcal{P}_1, \mathcal{P}_2|A^{\mathcal{O}}) \propto \sum_{\substack{\{m_{rs}\} \\ [m_{rs}=0,\ldots,n^0_{rs}]}} \frac{\prod_{r,s}(-1)^{m_{rs}}\binom{n^0_{rs}}{m}_{rs}}{\prod_r(n^1_r + m_r + 1)\prod_s(n^1_s + m_s + 1)}, \quad (3)$$

where the summation is over all possible values of each m_{rs} and, for clarity, we use the shorthand $r \equiv \alpha\beta$ and $s \equiv \gamma\delta$, $m_r \equiv \sum_s m_{rs}$ and $m_s \equiv \sum_r m_{rs}$ [45].

Given Eq. (3), which is the complete probabilistic description of the multilayer SBM, one could, in principle, find the partitions $\mathcal{P}_1$ and $\mathcal{P}_2$ that maximize $P_{\text{AND}}(\mathcal{P}_1, \mathcal{P}_2|A^{\mathcal{O}})$. If this were possible, one would be able to perfectly disentangle the two SBMs responsible for the observed links, even though the observation did not have explicit information about the layers. It would also be possible to compare regular SBMs to multilayer SBMs to determine if a multilayer model is more or less appropriate to describe a given network. Unfortunately, the expression above becomes numerically intractable even for a small number of groups and therefore one needs to make approximations that simplify the problem.

III. LINK RELIABILITY WITH APPROXIMATE MULTILAYER STOCHASTIC BLOCK MODELS

We propose an approximation that makes it possible to work with multilayer SBMs. We start by noting that any multilayer SBM can be represented as a single-layer SBM [Fig. 2(a)] [46]. In the single-layer SBM, each group comprises the nodes that belong to the same pair of groups α, γ in $\mathcal{P}_1$ and β, δ in $\mathcal{P}_2$ in the multilayer SBM (and only those); we call the single-layer partition the intersection partition. Moreover, if group r in the intersection partition corresponds to groups α in $\mathcal{P}_1$ and β in $\mathcal{P}_2$, and group s in the intersection partition corresponds to groups γ in $\mathcal{P}_1$ and δ in $\mathcal{P}_2$, then the probability of connection in the single-layer SBM is $q^{\text{AND}}_{rs} = q^1_{\alpha\gamma} q^2_{\beta\delta}$ (for simplicity, we again focus on the AND model and leave the OR model for the appendixes). This fully determines the single-layer SBM.

Here, we make the following approximation: we keep the information of the partitions $\mathcal{P}_1$ and $\mathcal{P}_2$ in the intersection partition, but consider that the matrix elements q^{AND}_{rs}, while each is the result of the product of two factors, are all independent of each other [see Fig. 2(b)]. Since this approximation is equivalent to integrating separately every term with a different $(\alpha, \beta, \gamma, \delta)$ combination in Eq. (2), it follows that the integrated likelihood depends exclusively on the intersection partition. In other words, within this approximation all pairs of partitions $(\mathcal{P}_1, \mathcal{P}_2)$ with the same intersection partition $\mathcal{P}_I$ are equally likely, and it is no longer possible to uniquely determine the multilayer SBM that best describes the observed topology.

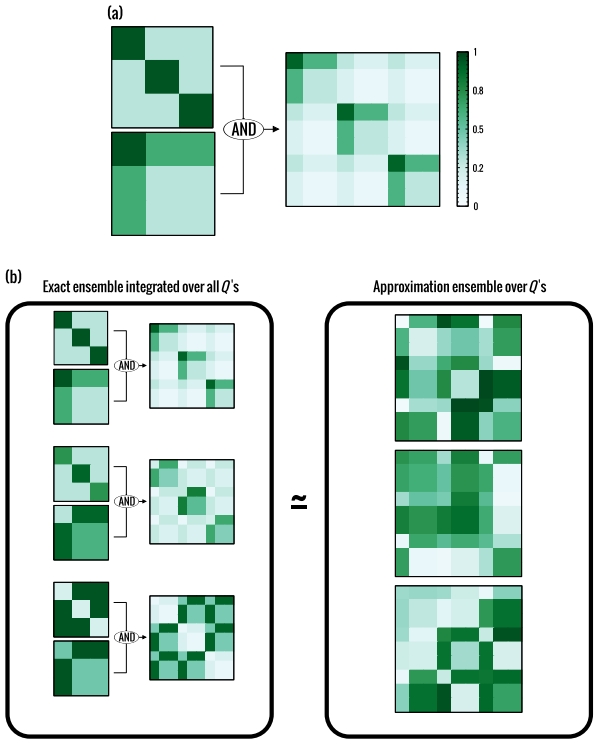

FIG. 2. Exact and approximate multilayer SBM ensembles. (a) Two independent single-layer SBMs aggregated using the AND mechanism. We represent each single-layer SBM by its node-to-node connection probability matrix (indicated in the shades of green shown in the color bar; note that node ordering is different in each SBM). The aggregation of the two layers can also be represented as a single-layer SBM, in which each group comprises the nodes that belong to the same pair of groups α in layer 1 and γ in layer 2; this is the intersection partition $\mathcal{P}_I$. Moreover, if group r in $\mathcal{P}_I$ corresponds to groups α in $\mathcal{P}_1$ and γ in $\mathcal{P}_2$, and group s in $\mathcal{P}_I$ corresponds to groups β in $\mathcal{P}_1$ and δ in $\mathcal{P}_2$, then the probability of connection in the single-layer SBM is $q^{\text{AND}}_{rs} = q^1_{\alpha\beta} q^2_{\gamma\delta}$. (b) For a fixed pair of partitions $\mathcal{P}_1$ and $\mathcal{P}_2$, we integrate over the ensemble of all possible probability matrices Q_1 and Q_2 [Eq. (3)]. For each pair (Q_1, Q_2), the resulting q^{AND}_{rs} are therefore correlated. In our approximation, we assume that the elements of the intersection q^{AND}_{rs} are randomly drawn and independent of each other.

Despite this limitation, our approximation still enables us to address the fundamental question of whether real-world networks are better described by single-layer or multilayer models. Specifically, in what follows we compare the predictive power of single-layer and multilayer SBMs in the problem of detecting missing and spurious links in noisy networks [27]. In fact, we argue that, if (approximate) multilayer SBMs yield better predictions on real networks, then there is evidence (supported by our results) to suggest that these networks are likely the outcome of multilayer processes (despite being observed as single-layer aggregates).

In the problem of assessing link reliability [27,47], the goal is to compute the probability $P(A_{ij}=1|A^{\mathcal{O}})$ that a link between nodes i and j truly exists ($A_{ij}=1$) given a noisy network observation $A^{\mathcal{O}}$, which contains false positives (spurious interactions that are reported but do not truly exist) and false negatives (missing interactions that truly exist but are not reported). We call the probability $R_{ij}=P(A_{ij}=1|A^{\mathcal{O}})$ the *reliability* of the link. In general, for any set $\mathcal{M}$ of models (single-layer SBMs, AND-multilayer SBMs, or OR-multilayer SBMs), the reliability is [27]

$$R_{ij}^{\mathcal{M}}=\frac{\int_{\mathcal{M}} dM P(A_{ij}=1|M)P(A^{\mathcal{O}}|M)P(M)}{Z}, \quad (4)$$

where Z is a normalization constant.

In the case of multilayer SBMs, the integral over the ensemble of models $\mathcal{M}$ requires (i) the integration over the connection probabilities Q_1 and Q_2 [akin to what we did to obtain Eq. (1)] and (ii) the sum over all pairs of partitions $\mathcal{P}_1$ and $\mathcal{P}_2$. Within our approximation, the first step can be carried out analytically, but the second cannot (see the appendixes). However, always within our approximation, one can exploit the fact that the integral in Eq. (4) depends exclusively on the intersection partition $\mathcal{P}_I$ and map the sum over pairs of partitions onto a sum over a single partition. By doing so, we obtain the following expression for the link reliability (see appendixes for the analogous expression for the OR model):

$$R_{ij}^{\mathrm{AND}}=\frac{1}{Z}\sum_{\mathcal{P}_I}\left(\frac{n^1_{\sigma_i\sigma_j}+1}{n_{\sigma_i\sigma_j}+2}\frac{\sum_{k=n^1_{\sigma_i\sigma_j}+2}^{n_{\sigma_i\sigma_j}+2}\frac{1}{k}}{\sum_{k=n^1_{\sigma_i\sigma_j}+1}^{n_{\sigma_i\sigma_j}+1}\frac{1}{k}}D(\mathcal{P}_I)e^{-\mathcal{H}(\mathcal{P}_I)}\right), \quad (5)$$

where the sum is over all possible intersection partitions (that is, all single-level partitions), $n^1_{\alpha\beta}$ is the number of links between groups α and β in the intersection partition, $n_{\alpha\beta}=n^0_{\alpha\beta}+n^1_{\alpha\beta}$ is the number of (possible links between) pairs of nodes in groups α and β, and $D(\mathcal{P}_I)$ the number of pairs $(\mathcal{P}_1,\mathcal{P}_2)$ that have the same intersection partition $\mathcal{P}_I$ (the degeneracy of partition $\mathcal{P}_I$; see the appendixes). The *energy* $\mathcal{H}$ is

$$\mathcal{H}(\mathcal{P}_I)=\sum_{\alpha\leq\beta}\left[\ln(n_{\alpha\beta}+1)+\ln\binom{n_{\alpha\beta}}{n^1_{\alpha\beta}}-\ln\left(\sum_{k=n^1_{\alpha\beta}+1}^{n_{\alpha\beta}+1}\frac{1}{k}\right)\right], \quad (6)$$

where the sum is over all distinct pairs of groups in $\mathcal{P}_I$.

As in Ref. [27], the expression for the link reliability [Eq. (5)] is analogous to an ensemble average of an observable in statistical mechanics, giving $\mathcal{H}(\mathcal{P}_I)$ the meaning of an energy associated to a specific intersection partition. We can use a Markov chain Monte Carlo algorithm to compute numerically R_{ij} (see Supplemental Material [48] for details) [49]. As it turns out, $\mathcal{H}(\mathcal{P}_I)$ is equal to the energy obtained assuming a single SBM [Eq. (S2), Ref. [27]] plus a term that accounts for the product of two probabilities that generate each element of the intersection probability matrix. In a Bayesian context, we can interpret this term and the degeneracy $D(\mathcal{P}_I)$ as nonuniform priors for the intersection partitions.

IV. VALIDATION OF LINK RELIABILITY ESTIMATION IN MODEL NETWORKS

Now that we are able to estimate link reliabilities using our approximation to two-layer (AND and OR) SBMs [Eq. (5)], as well as single-layer SBMs [27], we compare the performance of these approaches at detecting missing and spurious interactions. Our expectation is that if real-world networks are truly the result of the aggregation of multiple layers, then assuming a two-layer structure should result in a higher accuracy.

Note that, because single-layer and two-layer models are identical models with a different prior, one may expect that they perform equally well in large enough networks. This is because when one has infinite available information about the system, the prior has no effect on the inference and therefore single-layer and two-layer models should be equally accurate. While this is indeed the case for simple modular SBMs whose group sizes increase with network size (see Fig. S11 in Supplemental Material [48]), this is not necessarily the case for real-world networks. Indeed, real-world networks have very heterogeneous connectivity patterns, and groups can be arbitrarily small regardless of network size, which makes it impossible to gather infinite information about those groups. In that case, the choice of prior does affect the inference protocol so that we expect a difference in accuracy between single- and two-layer SBMs. As we show in the following, our results for all the real-world networks we study confirm that there are differences between predictions based on single-layer and two-layer models.

To identify the limits of detectability in terms of the choice of two-layer SBM model, we first construct a set of multilayer test networks that have a well-defined block structure in each of the two layers, and that are aggregated using the AND or OR models (see Supplemental Material and Fig. 3). We parametrize this ensemble of networks using two variables: (i) the low-to-high connectivity ratio λ and (ii) the average connectivity of nodes k. For a fixed value of k, we expect to obtain larger accuracies for the easy cases, that is, for networks with a more marked block structure (i.e., low values of λ).

We consider the predictive power of each of the approaches at detecting [27,47] (i) missing links (we remove a fraction f of the links and compute the fraction of times that a removed link has a higher reliability than a

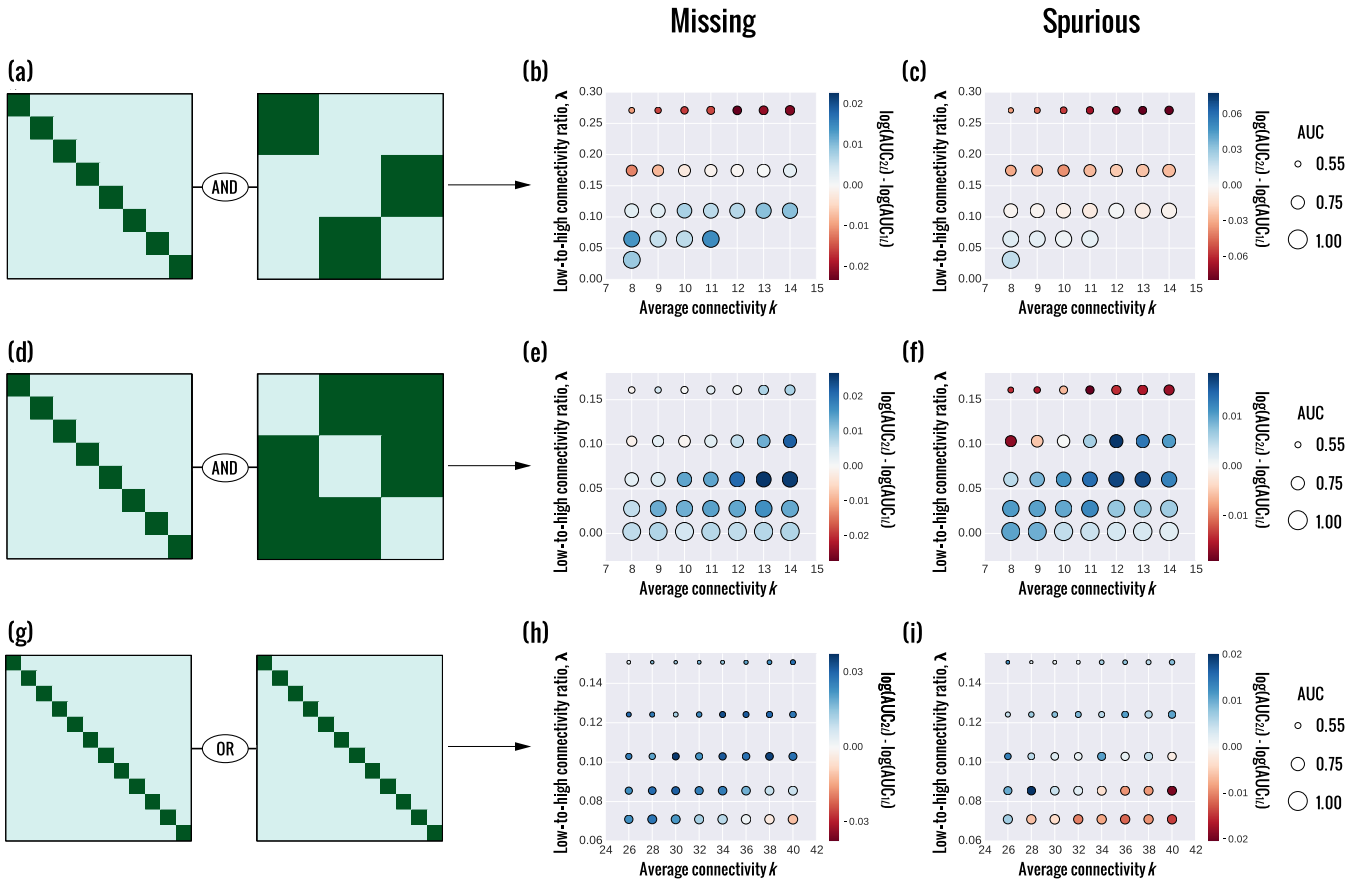

FIG. 3. Performance of missing and spurious link identification on synthetic aggregated two-layer networks. Each row shows results for the different sets of two-layer SBMs illustrated in (a),(d),(g). We consider networks of $N = 168$ (a),(d) and $N = 240$ (g) nodes divided into uniform groups in each layer. In the connection probability matrices, dark green represents a high connection probability p_h and light green a low connection probability p_l. We generate synthetic networks varying two parameters: the low-to-high connectivity ratio $\lambda = p_l/p_h < 1$ and the average connectivity k (see Supplemental Material [48]). To compare the performance of the different approaches at detecting missing links (b),(e),(h), we randomly remove a fraction $f = 0.25$ of the links (false negatives) from the real network and calculate the reliability of each unobserved link. Then we calculate the AUC statistic; that is, we rank the links by decreasing reliability and calculate how often a removed link (false negative) has a higher reliability than a link that does not exist in the original network (true negative). Analogously, to detect spurious links (c),(f),(i), we randomly add a fraction $f = 0.25$ of links (false positives), calculate the reliability of the observed links, and calculate how often an added link (false positive) has a lower reliability than a link that exists in the original network (true positive). For each pair of parameter values, we generate 30 different synthetic networks. We compare the average performance (AUC) at detecting missing links (b),(e),(h) and spurious links (c),(f),(i) of the approximate multilayer SBM approach, AUC_{2L}, against that of the single-layer SBM approach, AUC_{1L}. The size of the circles represents the AUC_{2L} of the multilayer approach. The color of the circles represents the logarithm of the ratio $\mathrm{AUC}_{2L}/\mathrm{AUC}_{1L}$, so that blue circles correspond to instances where the multilayer approach outperforms the single-layer approach, and conversely for red circles. [See Supplemental Material [48] for results for other values of f (fraction of false negatives/false positives) and for synthetic networks generated for different numbers of nodes and/or groups.]

link not present in the original network, that is, the AUC statistic) and (ii) spurious links (we add a fraction f of links and compute the fraction of times that an added link has a lower reliability than a link present in the original network, that is, the AUC statistic).

For AND networks [Figs. 3(a)–3(f) and Supplemental Material [48]) we find that, for the detection of both missing and spurious links, the two-layer approach outperforms the single-layer approach, especially (i) when the number of distinct node groups in the intersection partition and the connectivity grow and (ii) for small or moderate noise levels (fraction of removed or added links). Only when the structure of the blocks becomes very blurry do we observe that the single-layer approach works better (but in this region both approaches do, in fact, work poorly).

For OR networks [Figs. 3(g)–3(i) and Supplemental Material [48]), the two-layer approach again outperforms its single-layer counterpart in most situations. In this case, however, the largest improvements in performance happen for the hard cases (low accuracy values) with lower connectivity.

Note that in the OR model the aggregated network is denser than each of the layers, whereas in the AND model the aggregate is sparser than each of the original layers. For this reason, we expect the AND model to produce better results in real-world networks, which are sparse. In fact, we should expect the OR model to produce better results only for networks obtained from our ensemble of OR two-layer stochastic block models, that is, networks obtained from an OR aggregation of SBMs with independent and uniformly distributed probabilities of connection between pairs of groups (according to our prior). Our results for such an ensemble of networks confirm that this is the case (Fig. S12 in the Supplemental Material [48]).

V. MULTILAYER STOCHASTIC BLOCK MODELS ARE MORE PREDICTIVE FOR REAL NETWORKS

After showing that our approach is indeed more appropriate for model multilayer networks, we consider a real multilayer protein-protein interaction network of yeast *Saccharomyces cerevisiae*. In particular, we consider two types of interactions reported in the BioGRID database [50]: those detected using "two-hybrid" experiments and those obtained using "affinity-capture Western" experiments. We aggregate the two layers using the AND mechanism; that is, we build an aggregate network comprising the interactions that are detected by both types of experiments, and only those. As we show in Fig. 4, the multilayer model is again more accurate than the single-layer model at detecting missing and, especially, spurious interactions.

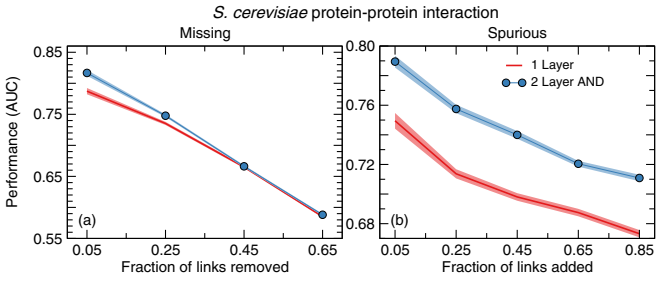

FIG. 4. Performance of missing and spurious link identification on a real multilayer network. We show the accuracy at detecting (a) missing and (b) spurious interactions in the protein-protein interaction network of S. cerevisiae, obtained from the BioGRID database [50], by aggregating two layers as described in the text. To compare the performance of the different approaches at detecting missing links (a), we randomly remove a fraction of the links (false negatives) from the real network and calculate the reliability of each unobserved link. Then we calculate the AUC statistic; that is, we rank the links by decreasing reliability and calculate how often a removed link (false negative) has a higher reliability than a link that does not exist in the original network (true negative). Analogously, to detect spurious links (b), we randomly add a fraction of links (false positives), calculate the reliability of the observed links, and calculate how often an added link (false positive) has a lower reliability than a link that exists in the original network (true positive).

Finally, we turn to the question of whether real networks that are observed as single-layer networks are, in fact, better described as aggregates of multiple layers. Specifically, we compare the performance of the single-layer and multilayer approaches on eight real-world networks (Fig. 5 and also Fig. S10 in the Supplemental Material [48]): (i) the air transportation network in Eastern Europe [51], (ii) the

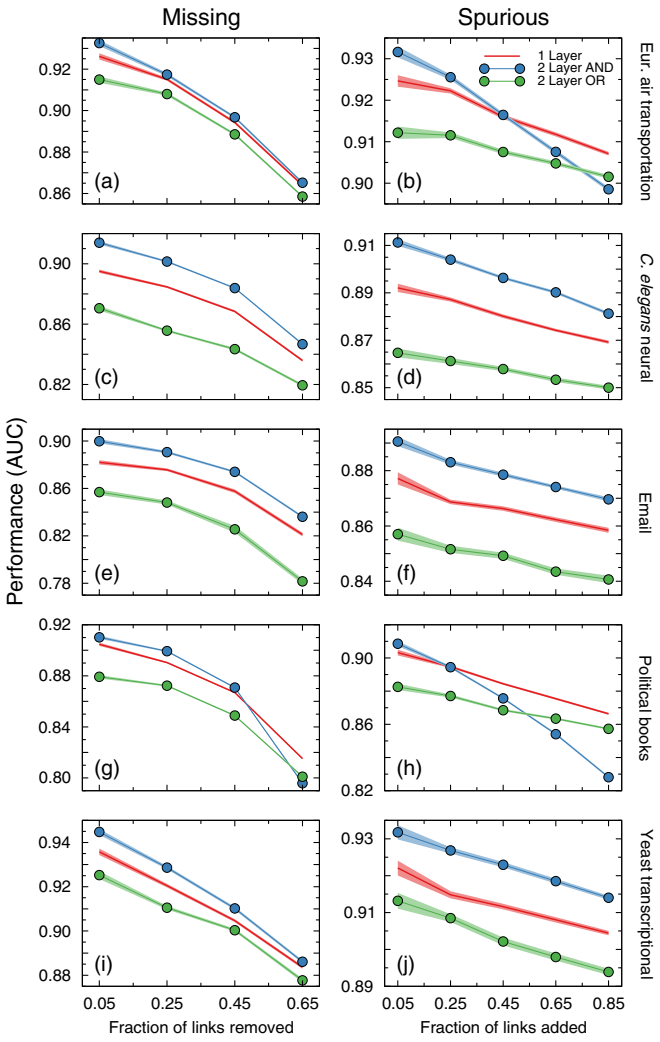

FIG. 5. Performance of missing and spurious link identification on real aggregated networks. We proceed as in Fig. 4 to compare the performance of the different approaches at detecting missing links (a),(c),(e),(g),(i) and spurious links (b),(d),(f),(h),(j). We show results for five real-world networks (see Supplemental Material for results on other networks [48]): (a),(b) the air transportation network in Eastern Europe [51], (c),(d) the neural network of *C. elegans* [52], (e),(f) the Email network within an organization [53], (g),(h) the network of books about U.S. politics in 2004 elections [54], and (i),(j) the transcriptional regulation network of yeast *S. cerevisiae* [55]. Red lines represent the AUC_{1L} obtained with single-layer SBMs, blue circles correspond to AND two-layer SBMs, and green circles to OR two-layer SBMs.

neural network of *Caenorhabditis elegans* [52], (iii) the Email network within a university [53], (iv) the network of frequent copurchasing of books about U.S. politics sold by the online bookseller Amazon during the 2004 presidential elections [54], (v) the transcriptional regulation network of yeast *S. cerevisiae* [55], (vi) the air transportation network in the U.S. [56], (vii) the collaboration network of jazz musicians, where two musicians are connected if they have played in the same band [57], and (viii) the network of American football games between colleges during regular season in the fall of 2000 [58].

Our results in Fig. 5 and also Fig. S10 of the Supplemental Material [48] show that the two-layer AND model provides a better description of these real-world networks since both missing and spurious interactions are consistently more accurately detected by the multilayer SBM approach, especially for low observational noise.

As mentioned earlier, a comparison of the two-layer approximation in Eq. (5) and the single-layer model in Ref. [27] shows that the two-layer model differs from the one-layer model in two ways. First, the AND model generates sparser networks than the single-layer model. Second, the two-layer model includes a degeneracy factor $D(\mathcal{P}_1)$ that favors partitions with a *larger* number of groups than the single-layer model (Table S3 in Supplemental Material [48]). Our results (Supplemental Material Fig. S10 [48]) show that neither of the two factors alone is responsible for the improvement in accuracy we observe. In particular, we show that if we add the degeneracy factor to the single-layer model, we already improve the accuracy at detecting missing and spurious links in most cases. From our results, it follows that sampling from partitions with a larger number of groups provides better models for real-world networks. This may seem counterintuitive, since one may expect a better model to have a lower number of parameters (groups in our case). However, because we expect the intersection block model resulting from a layer aggregation process to have a larger number of groups than the block models for each of the layers, this observation further reinforces our hypothesis that most real-world networks are in fact the result of an aggregation process.

VI. QUANTIFICATION OF THE PREFERENCE FOR MULTILAYER MODELS

Our results demonstrate that the two-layer stochastic block model (with AND aggregation) is more predictive for real-world complex networks, thus suggesting that real-world complex networks may be the result of the projection of several layers onto a single aggregate observation. To further quantify to what extent a two-layer model provides a better description of real-world networks than a single-layer model, we use Markov chain Monte Carlo sampling to compute the Bayes factor K of the models [59] (see Supplemental Material [48]), which is defined as

$$K = \frac{p(A^{\mathcal{O}}|M_2)}{p(A^{\mathcal{O}}|M_1)}. \tag{7}$$

Here, M_2 and M_1 are the two-layer AND SBM and the single-layer SBM, respectively, and the value of K represents the extent to which an observed network A^O supports the claim that the "true" model is the two-layer versus the single-layer model (if $K > 1$, model M_2 is better supported by the data under consideration than model M_1, and vice versa.)

Figure 6 shows that, for all the real-world networks we consider, the Bayes factor is larger than 1. Using the qualitative scale proposed by Kass and Raftery to map K values to human perception of evidence strength [60], we conclude that there is "very strong evidence" supporting the two-layer model for most of the networks; for the *S. cerevisiae* protein-protein interaction network and for the networks of political books the evidence is "strong." Importantly, Fig. 6 also shows that the preference for the two-layer model cannot be solely attributed to the sparsity induced by the AND aggregation. Indeed, simply adding to the single-layer model the degeneracy factor (which, as discussed above, introduces a preference for larger numbers of groups) also results in a model that is better supported by the data than the single-layer model, in all but one of the

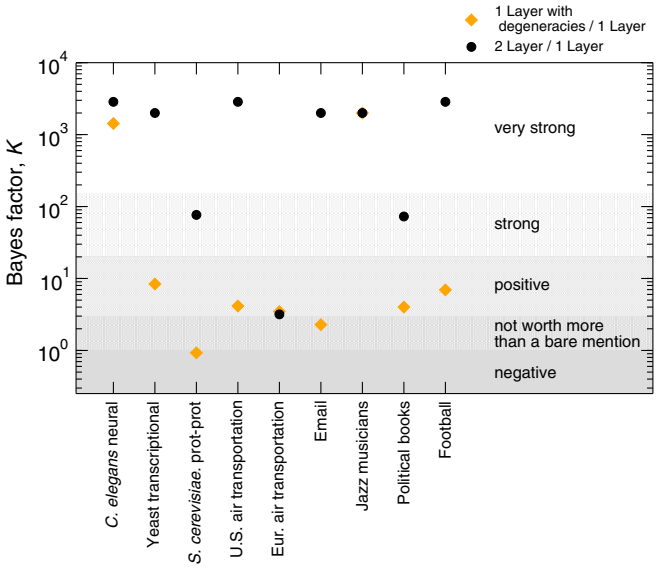

FIG. 6. Model comparison using Bayes factors. We show the Bayes factor K [Eq. (7)] for all the real-world network examples we analyze (see text). Black dots show results for the Bayes factor considering the two-layer AND SBM and the single-layer SBM. Orange diamonds show the Bayes factor considering the single-layer SBM with a degeneracy factor and the single-layer SBM. The regions in different shades of gray indicate the qualitative scale introduced by Kass and Raftery to map K values to human perceptions of strength of evidence [60] ("negative," $K < 1$; "not worth more than a bare mention," $1 < K < 3$; "positive," $3 < K < 20$; "strong," $20 < K < 150$; "very strong," $K > 150$).

networks we consider (the protein-protein interaction network of *S. cerevisiae*).

Interestingly, we find that there are some discrepancies between the evidence we find for the two air transportation networks we consider. While for U.S. air transportation we find very strong evidence for the two-layer model, we find barely any positive evidence for European air transportation. From previous analyses, we know that air transportation networks have a very strong modular component driven by geopolitical factors [51,61]; nonetheless, there is arguably a second layer that may arise from the distinction between international hubs that connect to one another and local aiports that connect to hubs. Our results suggest that while U.S. air transportation shows strong evidence for those two layers, for the European air transportation network (which is smaller and has a much lower density, and where geography may play a stronger role because of the presence of political borders) the evidence of the two layers is less conclusive.

VII. DISCUSSION

We introduce the family of multilayer SBMs, which generalizes single-layer SBMs to situations where links arise in different layers and are aggregated through different mechanisms. We also give the probabilistically complete solution to the problem of inferring the optimal multilayer SBM for a given aggregate network, and propose a tractable approximation that enables us to objectively address the question of whether an observed network is best described as the projection of multiple layers or as a single layer. Our results suggest that many real-world networks are indeed projections.

Although, as mentioned above, there have been proposals to extend the concept of modularity to multilayer networks [39], our approach represents a pioneering attempt to extend stochastic block models to multilayer systems. In this regard, it is important to stress that in this work we are concerned with the learning of multilayer models from aggregate networks where all information about the layers has been lost; in this sense, our work is different from previous attempts to do inference of stochastic block models on multigraphs where the layers themselves are observed [29].

Our work is also different from works on link prediction using latent feature models [40–42]. An important difference between latent feature approaches and ours is that the latent feature model considers that the probability of the existence of a link is a function of the weighted sum of the interactions at the different layers; therefore, the latent feature model does not allow a physical interpretation of what each layer is and of how layers are combined. All in all, latent feature models are very well suited for the inference of unobserved links, but due to the intricacies of the model and the difficulty to interpret its "parameters," it is not clear whether they are appropriate to address the question of whether a real network is really the outcome of processes occurring in different layers or not (and may also be prone to overfitting when observational data is noisy).

Our multilayer SBM is the simplest group-based multilayer model one can propose. Although our approach is not exempt of limitations (for example, it is computationally expensive and is therefore not suitable to handle extremely large networks), we believe that its detailed analysis will open the door to better understand the structure of real complex networks.

ACKNOWLEDGMENTS

We thank A. Aguilar-Mogas, A. Arenas, M. De Domenico, A. Godoy-Lorite, T. P. Peixoto, N. Rovira-Asenjo, O. Senan-Campos, and M. Tarrés-Deulofeu for helpful comments and discussions. This work was supported by a James S. McDonnell Foundation Research Award, Spanish Ministerio de Economía y Competitividad (MINECO) Grants No. FIS2013-47532-C3 and No. FIS2015-71563-ERC (R. G.), European Union Grant No. PIRG-GA-2010-277166 (R. G.), European Union Grant No. PIRG-GA-2010-268342 (M. S.-P.), and European Union FET Grant No. 317532 (MULTIPLEX).

APPENDIX A: CALCULATION OF $P_{\mathrm{AND}}(\mathcal{P}_1\mathcal{P}_2|A^{\mathcal{O}})$ FOR A TWO-LAYER SBM

For the AND model, we need to integrate Eq. (1) for $P_{\mathrm{AND}}(\mathcal{P}_1\mathcal{P}_2|A^{\mathcal{O}})$ over $q^1_{\alpha\beta}$ and $q^2_{\gamma\delta}$ assuming uniform priors. To simplify the notation, we introduce two indices r and s, so that $r \equiv \alpha\beta$ and $s \equiv \gamma\delta$, and we drop the reference to layer 1 and 2 so that $q_r \equiv q^1_{\alpha\beta}$ and $q_s \equiv q^2_{\gamma\delta}$. In order to perform the integration over q_r, for example, we note that all the terms that contain q_r have the following form:

$$q_r^{\sum_s n^1_{rs}} \prod_s (1 - q_r q_s)^{n^0_{rs}} = q_r^{n^1_r} \prod_s \sum_{m_{rs}=0,\ldots,n^0_{rs}} \binom{n^0_{rs}}{m_{rs}} (-)^{m_{rs}} (q_r q_s)^{m_{rs}}, \tag{A1}$$

where $n^1_r = \sum_s n^1_{rs}$. Then, for fixed values of $\{m_{rs}\}$, we have that the contribution to the likelihood factorizes for every q_r and q_s as follows:

$$\int DQ_r \int DQ_s \prod_{r,s} \binom{n^0_{rs}}{m_{rs}} (-)^{m_{rs}} \prod_r q_r^{n^1_r + m_r} \prod_s q_s^{n^1_s + m_s}, \tag{A2}$$

where $\int DQ_r = \prod_r (\int_0^1 dq_r)$ and $m_r \equiv \sum_s m_{rs}$.

Integrating out the q_r's and q_s's, we obtain for the likelihood the expression in Eq. (3).

APPENDIX B: OR COMBINATION OF LAYERS

For the OR model, one can obtain an expression for the likelihood by noticing that the OR models is an AND model for the no links, that is nonexisting edges between pairs of nodes. The likelihood of the observed topology $A^{\mathcal{O}}$ given the model M_{OR} assuming two layers is then

$$P(A^{\mathcal{O}}|M_{\mathrm{OR}}) = \prod_{\substack{[\alpha\le\beta]\\[\gamma\le\delta]}}[(1-q^a_{\alpha\beta})(1-q^b_{\gamma\delta})]^{n^0_{\alpha\beta\gamma\delta}} \times [1-(1-q^a_{\alpha\beta})(1-q^b_{\gamma\delta})]^{n^1_{\alpha\beta\gamma\delta}}, \quad \text{(B1)}$$

where all quantities have the same definition as in Eq. (2).

Following the same steps as in Appendix A, we obtain the following expression for $P_{\mathrm{OR}}(\mathcal{P}_1\mathcal{P}_2|A^{\mathcal{O}})$:

$$P_{\mathrm{OR}}(\mathcal{P}_1,\mathcal{P}_2|A^{\mathcal{O}}) \propto \sum_{\substack{\{m_{rs}\}\\[m_{rs}=0,\ldots,n^1_{rs}]}} \frac{\prod_{r,s}(-1)^{m_{rs}}\binom{n^1_{rs}}{m}_{rs}}{\prod_r(n^0_r+m_r+1)\prod_s(n^0_s+m_s+1)}, \quad \text{(B2)}$$

where, as in Appendix A, we use the notation $r \equiv \alpha\beta$ and $s \equiv \gamma\delta$, and all the quantities have already been defined in Appendix A.

Finally, one can then compute the reliability for an OR combination of two layers as

$$R^{\mathrm{OR}}_{ij} = 1 - \frac{1}{Z}\sum_{\mathcal{P}_I}\left(\frac{n^0_{\sigma_i\sigma_j}+1}{n_{\sigma_i\sigma_j}+2}\frac{\sum_{k=n^0_{\sigma_i\sigma_j}+2}^{n_{\sigma_i\sigma_j}+2}\frac{1}{k}}{\sum_{k=n^0_{\sigma_i\sigma_j}+1}^{n_{\sigma_i\sigma_j}+1}\frac{1}{k}}D(\mathcal{P}_I)e^{-\mathcal{H}(\mathcal{P}_I)}\right), \quad \text{(B3)}$$

$$\mathcal{H}(\mathcal{P}_I) = \sum_{\alpha\le\beta\in\mathcal{P}_I}\left[\ln(n_{\alpha\beta}+1)+\ln\binom{n_{\alpha\beta}}{n^0_{\alpha\beta}}-\ln\left(\sum_{k=n^0_{\alpha\beta}+1}^{n_{\alpha\beta}+1}\frac{1}{k}\right)\right], \quad \text{(B4)}$$

where, as before, the sum is over all possible (intersection) partitions, Z is a normalization constant, and $D(\mathcal{P}_I)$ is the number of pairs of partitions that have the same intersection. In Eq. (B4), the sum is over all distinct pairs of blocks within a fixed partition, $n^1_{\alpha\beta} = \sum_{i\le j}A_{ij}\delta_{\sigma_i\alpha}\delta_{\sigma_j\beta}$, $n_{\alpha\beta} = \sum_{i\le j}\delta_{\sigma_i\alpha}\delta_{\sigma_j\beta}$, $n^0_{\alpha\beta} = n_{\alpha\beta} - n^1_{\alpha\beta}$, and σ_i stands for the block to which node i belongs.

APPENDIX C: COMPUTATION OF DEGENERACIES

Our goal is to compute the number $D(\mathcal{P}_I)$ of pairs $(\mathcal{P}_1,\mathcal{P}_2)$ that have the same intersection partition $\mathcal{P}_I$, that is, the cardinality of the set $\{(\mathcal{P}_i,\mathcal{P}_j)|\mathcal{P}_i\cap\mathcal{P}_j = \mathcal{P}_I\}$. We start by noting that a specific $\mathcal{P}_I$ consists of n groups of nodes that we call "elements"; we make explicit the number of such elements in an intersection partition and write $\mathcal{P}^n_I = [E_1][E_2][E_3]\ldots[E_n]$. By the definition of intersection partition, we have that (i) all the nodes within an element must belong to the same group in both partitions $\mathcal{P}_1$ and $\mathcal{P}_2$ (otherwise, they would not belong to the same element) and (ii) two elements cannot belong to the same group in both $\mathcal{P}_1$ and $\mathcal{P}_2$ (otherwise, they would be a single element). We compute the degeneracy in two steps (see Supplemental Material for details [48]): (1) We compute all the possible unique partitions $\mathcal{P}_1$ combining the elements in $\mathcal{P}^n_I$, group them in *classes* according to the numbers of elements combined, and compute the multiplicity associated to each class, and (2) for each class, we compute all the possible partitions $\mathcal{P}_2$ that result in a specific intersection $\mathcal{P}^n_I$.

[1] A.-L. Barabási and Z. N. Oltvai, *Network Biology: Understanding the Cell's Functional Organization*, Nat. Rev. Genet. **5**, 101 (2004).

[2] E. Bullmore and O. Sporns, *Complex Brain Networks: Graph Theoretical Analysis of Structural and Functional Systems*, Nat. Rev. Neurosci. **10**, 186 (2009).

[3] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, *Network Medicine: A Network-Based Approach to Human Disease*, Nat. Rev. Genet. **12**, 56 (2011).

[4] P. Csermely, T. Korcsmáros, H. J. M. Kiss, G. London, and R. Nussinov, *Structure and Dynamics of Molecular Networks: A Novel Paradigm of Drug Discovery: A Comprehensive Review*, Pharmacol. Therapeut. **138**, 333 (2013).

[5] R. M. Thompson, U. Brose, J. A. Dunne, R. O. Hall, S. Hladyz, R. L. Kitching, N. D. Martinez, H. Rantala, T. N. Romanuk, D. B. Stouffer, and J. M. Tylianakis, *Food Webs: Reconciling the Structure and Function of Biodiversity*, Trends Ecol. Evol. **27**, 689 (2012).

[6] R. P. Rohr, S. Saavedra, and J. Bascompte, *On the Structural Stability of Mutualistic Systems*, Science **345**, 1253497 (2014).

[7] F. Schweitzer, G. Fagiolo, D. Sornette, F. Vega-Redondo, A. Vespignani, and D. R. White, *Economic Networks: The New Challenges*, Science **325**, 422 (2009).

[8] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca, *Network Analysis in the Social Sciences*, Science **323**, 892 (2009).

[9] M. E. J. Newman, *Communities, Modules and Large-Scale Structure in Networks*, Nat. Phys. **8**, 25 (2011).

[10] R. Guimerà and L. A. N. Amaral, *Functional Cartography of Complex Metabolic Networks*, Nature (London) **433**, 895 (2005).

[11] A. Arenas, A. Díaz-Guilera, and C. J. Pérez-Vicente, *Synchronization Reveals Topological Scales in Complex Networks*, Phys. Rev. Lett. **96**, 114102 (2006).

[12] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, *Classes of Complex Networks Defined by Role-to-Role Connectivity Profiles*, Nat. Phys. **3**, 63 (2007).

[13] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, *Link Communities Reveal Multiscale Complexity in Networks*, Nature (London) **466**, 761 (2010).

[14] S. Fortunato, *Community Detection in Graphs*, Phys. Rep. **486**, 75 (2010).

[15] H. C. White, S. A. Boorman, and R. L. Breiger, *Social Structure from Multiple Networks. I. Blockmodels of Roles and Positions*, Am. J. Sociology **81**, 730 (1976).

[16] P. W. Holland, K. B. Laskey, and S. Leinhardt, *Stochastic Blockmodels: First Steps*, Soc. Networks **5**, 109 (1983).

[17] K. Nowicki and T. A. B. Snijders, *Estimation and Prediction for Stochastic Blockstructures*, J. Am. Stat. Assoc. **96**, 1077 (2001).

[18] B. Karrer and M. E. J. Newman, *Stochastic Blockmodels and Community Structure in Networks*, Phys. Rev. E **83**, 016107 (2011).

[19] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, *Inference and Phase Transitions in the Detection of Modules in Sparse Networks*, Phys. Rev. Lett. **107**, 065701 (2011).

[20] M. N. Schmidt and M. Mørup, *Non-Parametric Bayesian Modeling of Complex Networks. An Introduction*, IEEE Signal Process. Mag. **30**, 110 (2013).

[21] T. P. Peixoto, *Parsimonious Module Inference in Large Networks*, Phys. Rev. Lett. **110**, 148701 (2013).

[22] T. P. Peixoto, *Hierarchical Block Structures and High-Resolution Model Selection in Large Networks*, Phys. Rev. X **4**, 011047 (2014).

[23] T. P. Peixoto, *Efficient Monte Carlo and Greedy Heuristic for the Inference of Stochastic Block Models*, Phys. Rev. E **89**, 012804 (2014).

[24] D. B. Larremore, A. Clauset, and A. Z. Jacobs, *Efficiently Inferring Community Structure in Bipartite Networks*, Phys. Rev. E **90**, 012805 (2014).

[25] C. Aicher, A. Z. Jacobs, and A. Clauset, *Learning Latent Block Structure in Weighted Networks*, J. Complex Netw. **3**, 221 (2015).

[26] X. Yan, C. Shalizi, J. E. Jensen, F. Krzakala, C. Moore, L. Zdeborová, P. Zhang, and Y. Zhu, *Model Selection for Degree-Corrected Block Models*, J. Stat. Mech. (2014) P05007.

[27] R. Guimerà and M. Sales-Pardo, *Missing and Spurious Interactions and the Reconstruction of Complex Networks*, Proc. Natl. Acad. Sci. U.S.A. **106**, 22073 (2009).

[28] R. Guimerà and M. Sales-Pardo, *Justice Blocks and Predictability of U.S. Supreme Court Votes*, PLoS One **6**, e27188 (2011).

[29] R. Guimerà, A. Llorente, E. Moro, and M. Sales-Pardo, *Predicting Human Preferences Using the Block Structure of Complex Social Networks*, PLoS One **7**, e44620 (2012).

[30] N. Rovira-Asenjo, T. Gumí, M. Sales-Pardo, and R. Guimerà, *Predicting Future Conflict between Team-Members with Parameter-Free Models of Social Networks*, Sci. Rep. **3**, 1999 (2013).

[31] R. Guimerà and M. Sales-Pardo, *A Network Inference Method for Large-Scale Unsupervised Identification of Novel Drug-Drug Interactions*, PLoS Comput. Biol. **9**, e1003374 (2013).

[32] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, *Multilayer Networks*, J. Complex Netw. **2**, 203 (2014).

[33] R. G. Morris and M. Barthelemy, *Transport on Coupled Spatial Networks*, Phys. Rev. Lett. **109**, 128703 (2012).

[34] F. Radicchi and A. Arenas, *Abrupt Transition in the Structural Formation of Interconnected Networks*, Nat. Phys. **9**, 717 (2013).

[35] 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).

[36] M. De Domenico, A. Solé-Ribalta, E. Cozzo, M. Kivelä, Y. Moreno, M. A. Porter, S. Gómez, and A. Arenas, *Mathematical Formulation of Multilayer Networks*, Phys. Rev. X **3**, 041022 (2013).

[37] 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).

[38] J. Sanz, C.-Y. Xia, S. Meloni, and Y. Moreno, *Dynamics of Interacting Diseases*, Phys. Rev. X **4**, 041005 (2014).

[39] 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).

[40] K. Miller, M. I. Jordan, and T. L. Griffiths, in *Advances in Neural Information Processing Systems 22*, edited by Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Curran Associates, Inc., Redd Hook, NY, 2009), pp. 1276–1284.

[41] K. Palla, D. Knowles, and Z. Ghahramani, in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)* (Omnipress, Madison, WI, 2012), pp. 1607–1614.

[42] M. Kim and J. Leskovec, in *Advances in Neural Information Processing Systems 26*, edited by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Curran Associates, Inc., Redd Hook, NY, 2013), pp. 1385–1393.

[43] L. Weng, F. Menczer, and Y.-Y. Ahn, *Virality Prediction and Community Structure in Social Networks*, Sci. Rep. **3**, 2522 (2013).

[44] Other possibilities include choosing nonuniform priors for the connection probabilities [20–22] or different priors for the partitions [19–23,26].

[45] Note that it is straightforward to apply the same formalism to directed networks by considering a bipartite graph of nodes with incoming and outgoing connections. In this case, for each layer we would have a SBM with two sets of block partitions, one for nodes with outgoing connections and one for nodes with incoming connections, and a nonsymmetric connection probability matrix (see Ref. [30]).

[46] The reverse is also true, so the possible network models one can generate with single-layer SBMs and multilayer SBMs are, in fact, identical. For instance, a specific one-layer SBM is equivalent to a two-layer AND model in which one of the layers has a single group and connection probability matrix with all entries equal to 1 and another layer equal to the single-layer SBM. However, it is important to note that each of them gives different weights to different models, so that a model that is relatively probable in the multilayer SBM family might be relatively rare in the single-layer SBM family, and vice versa.

[47] A. Clauset, C. Moore, and M. E. J. Newman, *Hierarchical Structure and the Prediction of Missing Links in Networks*, Nature (London) **453**, 98 (2008).

[48] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevX.6.011036 for detailed

description of the methodologies and calculations mentioned in the main text. Supplementary results not shown in the figures of the main text.

[49] Note that, although Eqs. (5) and (6) are the exact solution to the link inference problem with approximate multilayer stochastic block models, there is no mathematical guarantee that, in a finite amount of time, the Markov chain will sample the space of node partitions with the desired probabilities. In particular, the energy landscape may be rugged and the chain may get trapped in some region. However, we have performed equilibration tests that suggest that the chain is, indeed, sampling the space correctly.

[50] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, *BioGRID: A General Repository for Interaction Datasets*, Nucleic Acids Res. **34**, D535 (2006).

[51] R. Guimerà, S. Mossa, A. Turtschi, and L. A. N. Amaral, *The Worldwide Air Transportation Network: Anomalous Centrality, Community Structure, and Cities' Global Roles*, Proc. Natl. Acad. Sci. U.S.A. **102**, 7794 (2005).

[52] J. G. White, E. Southgate, J. N. Thomson, and S. Brenner, *The Structure of the Nervous System of the Nematode C. elegans*, Phil. Trans. R. Soc. B **314**, 1 (1986).

[53] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, *Self-Similar Community Structure in a Network of Human Interactions*, Phys. Rev. E **68**, 065103 (2003).

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

[55] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, *Network Motifs: Simple Building Blocks of Complex Networks*, Science **298**, 824 (2002).

[56] V. Batagelj and A. Mrvar, http://vlado.fmf.uni-lj.si/pub/networks/data.

[57] P. M. Gleiser and L. Danon, *Community Structure in Jazz*, Adv. Compl. Syst. **6**, 565 (2003).

[58] M. Girvan and M. E. J. Newman, *Community Structure in Social and Biological Networks*, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).

[59] E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, England, 2003).

[60] R. E. Kass and A. E. Raftery, *Bayes Factors*, J. Am. Stat. Assoc. **90**, 773 (1995).

[61] M. Sales-Pardo, R. Guimerà, A. A. Moreira, and L. A. N. Amaral, *Extracting the Hierarchical Organization of Complex Systems*, Proc. Natl. Acad. Sci. U.S.A. **104**, 15224 (2007).