Multiple phases in modularity-based community detection

Christophe Schülke[*]
Université Paris Diderot, Sorbonne Paris Cité, 75205 Paris, France
and Dipartimento di Fisica, Università di Roma "La Sapienza," Piazzale Aldo Moro 2, 00185 Rome, Italy

Federico Ricci-Tersenghi[†]
Dipartimento di Fisica, INFN–Sezione di Roma 1, and CNR-NANOTEC, UOS di Roma, Università di Roma "La Sapienza," Piazzale Aldo Moro 2, 00185 Rome, Italy

(Received 16 June 2015; revised manuscript received 22 September 2015; published 7 October 2015)

Detecting communities in a network, based only on the adjacency matrix, is a problem of interest to several scientific disciplines. Recently, Zhang and Moore have introduced an algorithm [Proc. Natl. Acad. Sci. USA **111**, 18144 (2014)], called mod-bp, that avoids overfitting the data by optimizing a weighted average of modularity (a popular goodness-of-fit measure in community detection) and entropy (i.e., number of configurations with a given modularity). The adjustment of the relative weight, the "temperature" of the model, is crucial for getting a correct result from mod-bp. In this work we study the many phase transitions that mod-bp may undergo by changing the two parameters of the algorithm: the temperature T and the maximum number of groups q. We introduce a new set of order parameters that allow us to determine the actual number of groups $\hat{q}$, and we observe on both synthetic and real networks the existence of phases with any $\hat{q} \in \{1,q\}$, which were unknown before. We discuss how to interpret the results of mod-bp and how to make the optimal choice for the problem of detecting significant communities.

DOI: 10.1103/PhysRevE.92.042804 PACS number(s): 64.60.aq, 89.20.−a

I. INTRODUCTION

In community detection, the goal is to regroup nodes of an observed network into different groups (or communities) of nodes believed to be similar and thus to find a meaningful partition of the network. The assumption that this is possible comes from the hypothesis that the structure of the graph reflects hidden attributes of the nodes that therefore can be inferred. Though recent studies show that such an assumption does not hold in general for real networks [1], generative models with this property, such as the stochastic block model (SMB) [2], have proved the efficiency of community detection algorithms [3]. Different classes of community detection algorithm exist: Among the most popular approaches, algorithms relying on Bayesian inference fit the parameters of an assumed generative model to the observed network [3,4], while spectral algorithms find communities from the eigenvectors of a matrix based on the adjacency matrix of the network [5,6].

The hypothesis most commonly made is that of assortative networks, which means that nodes with the same hidden attributes are more likely to be linked than nodes with different attributes. Under this hypothesis of assortative networks, a popular measure of the goodness of a partition is the modularity, and therefore various community detection algorithms rely on modularity maximization [7–10]. Recently, the authors of Ref. [11] (called ZM hereafter) introduced such an algorithm that avoids the common pitfall of overfitting: Indeed, maximizing modularity predicts communities even in unstructured (i.e., random) networks. The only free parameters in the mod-bp algorithm are the number of groups q and a temperature-like parameter T. Three ranges of temperatures are identified that correspond to phases in which the algorithm has qualitatively different behaviors: At high temperatures no division in groups is found, at intermediate temperatures meaningful groups are found, and at low-enough temperatures the algorithm does not converge.

Very often the success of an algorithm that has been designed for solving a given problem (detecting communities in the present case) is strongly related to the structure of problem's solutions; an eventual phase transition, i.e., a drastic rearrangement in the solutions space, may have a direct impact on the algorithm's behavior. For this reason the study of possible different phases arising in a given problem is essential to understand also the behavior of algorithms. In this paper, we broaden the picture given in ZM by showing that there are in general more than three phases. We show that despite passing the number q of groups to the mod-bp algorithm, it can spontaneously return a partition with a smaller number of groups $\hat{q} < q$. We introduce a new set of order parameters that allows us to determine $\hat{q}$ and observe both on synthetic and real networks the existence of phases with different values of $\hat{q} \in \{1,q\}$.

We will use the following notations: N is the number of nodes in the network, $\mathcal{E}$ is the set of m undirected edges, and we write $\langle ij \rangle \in \mathcal{E}$ if an edge is present between nodes i and j. The degree d_i of a node is the number of edges that link node i to other nodes. A partition of the network is a set $\{t\}$, where $t_i \in \{1,q\}$ is the group to which node i belongs. q is the maximum number of groups.

The modularity of a partition $\{t\}$ is defined by [12]

$$Q(\{t\}) = \frac{1}{m}\left(\sum_{\langle ij\rangle\in\mathcal{E}} \delta_{t_i,t_j} - \sum_{\langle ij\rangle} \frac{d_i d_j}{2m}\delta_{t_i,t_j}\right), \qquad (1)$$

where δ is the Kronecker δ function. High values of modularity indicate that there are more edges between nodes of the same

[*]christophe.schulke@espci.fr
[†]federico.ricci@roma1.infn.it

group than between nodes of different groups: Thus, the higher the modularity, the better the partition. The advantage of modularity is that it makes no assumption on the way the network was generated, but only that it has an assortative structure. This encourages its use on real networks, in which the true generative process is generally unknown, and has led to several algorithms performing community detection by maximization of modularity [7–10].

One drawback of modularity is that finding the partition with highest modularity is a discrete combinatorial optimization problem [13], which becomes rapidly intractable as N increases, so effective heuristics have to be developed. Another drawback is that modularity maximization is prone to overfitting: It is possible to find high-modularity partitions even in Erdős-Renyi random graphs [14], although by construction they do not contain an underlying group strucure [15–17]. Finally, there exists a fundamental resolution limit [18] that prevents the recovery of small-sized groups.

ZM introduces a new community-detection algorithm based on modularity maximization, tackling the two first mentioned drawbacks and proposing a multiresolution strategy to overcome the third. The algorithm, called mod-bp, is scalable, i.e., is of polynomial complexity with respect to N, and the authors show that it does not overfit, in the sense that it does not return high-modularity partitions for Erdős-Renyi networks.

This is achieved by treating modularity maximization as a statistical physics problem with an energy

$$E(\{t\}) = -m\,Q(\{t\}) \tag{2}$$

at a finite temperature $T = \frac{1}{\beta}$. In this way, every partition $\{t\}$ is given a probability taken from the Gibbs distribution

$$P(\{t\}) = \frac{1}{Z} e^{-\beta E(\{t\})}, \tag{3}$$

where Z is the partition function

$$Z = \sum_{\{t\}} e^{-\beta E(\{t\})}. \tag{4}$$

To solve the problem of sampling from the Gibbs distribution (3), ZM proposes a belief propagation (BP) algorithm [19,20], in which so-called messages $\psi_t^{i\to k}$ are sent between all pairs of nodes $\langle ik\rangle$ for q different groups t. We refer the reader to ZM for a precise description of the algorithm. After convergence of the BP algorithm, marginals ψ_t^i are obtained from the messages. ψ_t^i represents the probability that node i belongs to group t, and the most-likely group for node i is therefore:

$$\hat{t}_i = \arg\max_t \psi_t^i. \tag{5}$$

Using this maximization, the maximum *a posteriori* modularity Q^{MAP} corresponding to the assignment $\{\hat{t}\}$ can be calculated as

$$Q^{\mathrm{MAP}} = Q(\{\hat{t}\}). \tag{6}$$

As the algorithm samples from the distribution (3), one can also define an average modularity Q^{MARG} that is calculated from the marginals instead of the most-likely partition and which is proportional to the average energy of the model,

$$Q^{\mathrm{MARG}} = \frac{1}{m}\frac{\partial}{\partial\beta}\left(\sum_i \log Z_i - \sum_{\langle ij\rangle\in\mathcal{E}} \log Z_{ij} + \frac{\beta}{4m}\sum_t \theta_t^2\right),$$

where Z_i and Z_{ij} are the normalizations of the marginals and the two-point correlation functions, respectively, and $\theta_t = \sum_{j=1}^N d_j \psi_t^j$.

While the problem of maximizing modularity is equivalent to finding the ground state of (2), sampling from (3) at a finite temperature corresponds to minimizing the corresponding free energy. This means taking into account not only the modularity but also the entropy, counting the number of partitions with a given modularity. In this way, instead of focusing on a single partition, mod-bp at finite T returns a partition that is a good consensus of the many existing high-modularity partitions, as advocated in Ref. [21].

II. PHASE TRANSITIONS

As in numerous statistical physics problems, (3) may lead to phase transitions at some given temperatures. Using modularity as an energy function is similar to studying a Potts-like statistical mechanics problem [22], for which Ref. [23] has shown that a phase transition is always present. ZM reports that temperature ranges define three different regimes of the algorithm. At very low temperatures, the system is in a spin-glass phase, in which the algorithm does not converge to a fixed point. At high temperature, the system is in a paramagnetic phase in which the fixed point is trivial and all nodes have an equal probability $1/q$ of belonging to any of the groups. In networks with statistically significant communities, there is an intermediate temperature range called the recovery phase, in which the algorithm converges to a nontrivial fixed point, from which group assignments can be obtained using (5).

Here we broaden this picture by showing that the recovery phase itself can be divided in up to $q-1$ phases, with $2 \leqslant \hat{q} \leqslant q$. Approaching the temperature separating two phases, there is an order parameter that becomes vanishingly small, and the number of iterations needed by the algorithm to reach the fixed point diverges.

A. Model-based critical temperatures

Modularity as a measure of goodness of a partition is particularity appealing for real networks, because it makes no assumption about an underlying model that generates the network. Though appealing, this absence of model is problematic when it comes to determining the best temperature at which to run mod-bp (i.e., there is no Bayes optimal temperature). In ZM two generative models are analyzed, allowing us to find two useful characteristic temperatures:

(1) In the configuration model, a network is built by randomly creating links between nodes of known degrees, until all nodes have the right number of neighbors. ZM shows that in this model, the phase transition between the spin-glass

and the paramagnetic phase takes place at

$$T^* = \left[\log\left(\frac{q}{\sqrt{c}-1}+1\right)\right]^{-1}, \tag{7}$$

where c is the average excess degree, calculated from the average degree $\langle d\rangle$ and the average squared degree $\langle d^2\rangle$ and given by

$$c = \frac{\langle d^2\rangle}{\langle d\rangle} - 1. \tag{8}$$

(2) In the stochastic block model (SBM) [2], the nodes are grouped into q^* equal-sized groups, and for each pair of nodes $\langle ij\rangle$, a link is created with probability p_{rs} if i belongs to group r and and j belongs to group s. In the most simple case, we take $p_{rs} = p_{\text{out}}$ if $r \neq s$ and $p_{rs} = p_{\text{in}}$ if $r = s$. One often considers networks with sparse connectivity, i.e., the average number of links between a node i from group r and all the nodes from group s, c_{rs}, does not grow with the size of the network. ZM shows that mod-bp is as successful as a Bayes-optimal algorithm [24] and that the transition between the paramagnetic phase and the recovery phase takes place at

$$T_R(\epsilon) = \left(\log\left\{\frac{q[1+(q-1)\epsilon]}{c(1-\epsilon)-[1+(q-1)\epsilon]}+1\right\}\right)^{-1}, \tag{9}$$

where $\epsilon = p_{\text{out}}/p_{\text{in}}$.

For real retworks, the stochastic block model is usually a bad model, and the recommendation of ZM is to run the algorithm at T^*, which seems to always lie inside of the recovery phase. We can also note that the $\epsilon \to 0$ limit of (9),

$$T_0 = \left[\log\left(\frac{q}{c-1}+1\right)\right]^{-1}, \tag{10}$$

is a useful upper bound for T. Indeed, above this temperature, the algorithm converges to the paramagnetic solution, even for networks composed of disconnected components, and is therefore useless.

B. Degenerate groups

In the paramagnetic phase, the marginal ψ_t^i of every node i of the network is equal to $\frac{1}{q}$ for all t, up to some minor fluctations due to the numerical precision of the machine or incomplete convergence of the algorithm. Due to those fluctuations, calculating a retrieval configuration with (5) is in general still possible and would lead to a very small but nonvanishing retrieval modularity Q^{MAP}.

However, the meaning of the paramagnetic phase is that all groups are strictly equivalent or degenerate, and therefore Q^{MAP} should be exactly zero. In order to obtain this, the algorithm has to check for degenerate groups before assigning a group to each node and assign the same "effective" group to nodes for which the maximization (5) leads to different, but actually degenerate, groups.

To check if groups are degenerate, we can look at the following distance between two groups k and l:

$$d_{kl} = \frac{1}{N}\sum_{i=1}^{N}\left(\psi_k^i - \psi_l^i\right)^2. \tag{11}$$

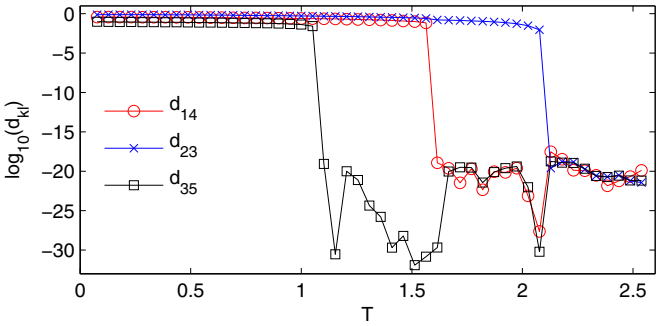

FIG. 1. (Color online) d_{12}, d_{23}, and d_{35} as a function of temperature. In order to follow the groups at different temperatures, the temperature is increased step by step, and the messages are initialized with the final values they reached at the last temperature. We see that the group distances d_{kl} are like order parameters undergoing a phase transition at different temperatures, where they drop by more than 10 orders of magnitude. Due to this phase transition, it is easy to choose a threshold d_{min} in Eq. (12). The data set is "political books," run with $q = 6$.

If d_{kl} is smaller than a choosen threshold d_{min}, then we can consider that group k and group l are degenerate and that they should not be distinguished.

An effective number of groups, $\hat{q}$, then can be defined as the number of *distinguishable* groups. We can define a mapping ϕ between the q groups used by the algorithm and the $\hat{q}$ distinguishable groups: For each group k, $\phi(k)$ is an integer between 1 and $\hat{q}$ representing one of the effective groups, and

$$\forall(k,l), \quad \phi(k) = \phi(l) \Leftrightarrow d_{kl} < d_{\text{min}}. \tag{12}$$

With this mapping, we replace the group assignment procedure (5) by

$$\hat{t}_i = \phi\left(\arg\max_t \psi_t^i\right). \tag{13}$$

With this assignment procedure, Q^{MAP} is strictly zero in the paramagnetic phase, because all nodes belong to the same group.

Figure 1 shows that choosing a threshold d_{min} is meaningful because d_{kl} undergoes a phase transition at which it sharply drops of several orders of magnitude.

Interestingly, group degeneracy is observed not only in the paramagnetic phase but also inside the retrieval phase. In that case, not all groups are degenerate but only a subset of them. Figure 2 shows this for the popular network "political books" [25], on which mod-bp was run at different temperatures.

III. EXISTING DOMAINS OF PHASES

Thanks to the group assignment procedure in Eq. (13), up to $q+1$ phases can exist for any network on which mod-bp is run with q groups: one for each $\hat{q} \in \{1,q\}$, plus a spin-glass phase. Figure 3 shows this for the network "political books." On this network, several phases coexist at low temperature, whereas for higher temperatures, the phases exist in well-separated temperature intervals. In the latter case, we can define a "critical" temperature T_k, separating the phase with $\hat{q} = k$ from the one with $\hat{q} = k+1$. As can be seen on Fig. 3, the number

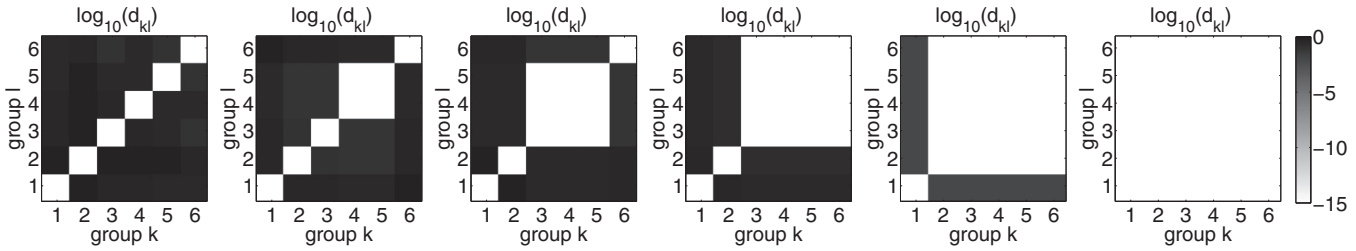

FIG. 2. Matrices of distances between groups for different temperatures. The data set is "political books," and the algorithm is run with $q = 6$ for $T = 0.26, 0.9, 1.0, 1.28, 2.1, 2.3$ (from left to right). We observe the formation of a growing cluster of groups that are equivalent, allowing us to define a number of effective groups $\hat{q}$ that varies from 6 at low temperature (left) to 1 in the paramagnetic phase (right). Note that the area of the squares is not related to the number of nodes contained in the groups.

of iterations needed for mod-bp to converge greatly increases around these critical temperatures. As noted before, T_0 is a good reference temperature, and normalizing all temperatures by T_0 is a good way to compare critical temperatures T_k for the same network with different q values and for comparing different networks.

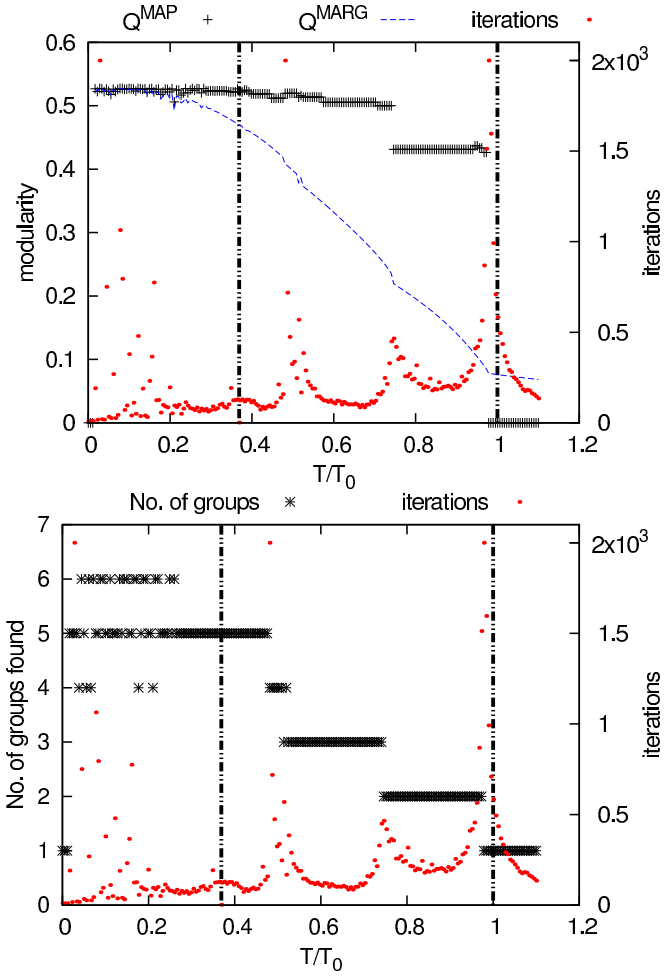

FIG. 3. (Color online) Modularities and numbers of effective groups $\hat{q}$ obtained by sweeping a temperature range from 0 to 1.2 T_0 on the data set "political books" with $q = 6$. The vertical lines indicate the positions of T^* (left) and T_0 (right). Above T^*, the changes in $\hat{q}$ define quite homogeneous phases, separated by sharp transitions, where the number of iterations necessary to reach convergence increases greatly. Below $T/T_0 \approx 0.4$, the phase is not homogeneous: Depending on the starting conditions, $\hat{q}$ can be 4, 5, or 6. Note that Q^{MAP} increases only minimally when $\hat{q}$ exceeds 3, which agrees with the fact that $q^* = 3$.

A. Location of critical points T_k

In some cases, a subset of n critical temperatures T_k can be degenerate, in which case there is a phase transition between a phase with $\hat{q} = k$ and a phase with $\hat{q} = k + n$. For instance, this is the case in networks generated by the stochastic block model with the same in-connectivity p_{in} inside each of the q^* groups (Fig. 4, top). This agrees with the description of the three phases given in ZM.

In contrast, in networks generated by the SBM with $p_{rr} \neq p_{tt}$, if $r \neq t$, then the degeneracy is lifted (Fig. 4, bottom). The figure also shows that, starting above T_0 (i.e., in the paramagnetic phase) and lowering the temperature, the groups are inferred in order of their strength.

To show this, we use the recall score for different groups, which allows us to see if one of the inferred groups corresponds well to a given real group. To quantify the similarity between a real group G and an inferred group $\hat{G}_i$ that are not necessarily of the same size, we can use the Jaccard score [1], which is defined by:

$$J(G,\hat{G}_i) = \frac{|G \cap \hat{G}_i|}{|G \cup \hat{G}_i|}. \tag{14}$$

The recall score is the maximum of the Jaccard score:

$$R(G) = \max_i J(G,\hat{G}_i). \tag{15}$$

A recall score close to 1 means that one of the inferred groups $\hat{G}_i$ is almost identical to group G. Figure 4 (bottom) therefore shows that around $T/T_0 = 0.8$, the group with the biggest in-connectivity is nearly exactly recovered by one of the groups returned by the algorithm, whereas the two groups with lower in-connectivity are not. Only by further lowering the temperature, when $\hat{q} = 3$, are all the groups correctly inferred.

B. Running mod-bp with $q \neq q^*$

On networks generated with the SBM, the real number of groups q^* is known, and it is thus interesting to look at what happens when mod-bp is run with $q \neq q^*$. The behavior for $q = q^*$ is described in ZM and in Fig. 4. If $q < q^*$, then mod-bp cannot return the right number of groups and will

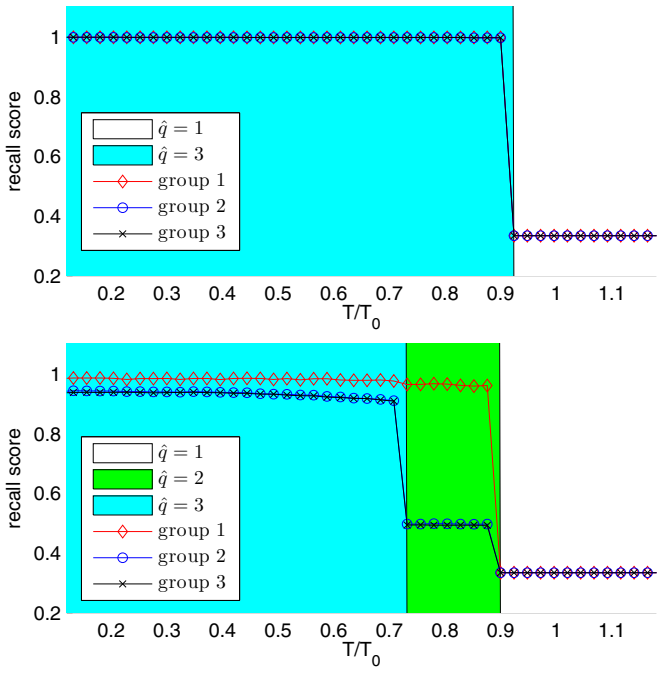

FIG. 4. (Color online) Degeneracy of T_ks on networks generated by the SBM with $N = 5000$, $q^* = 3$ and $c_{\text{out}} = 2$. mod-bp was run with $q = 3$. Top: All 3 groups have the same in-connectivity $c_{\text{in}} = 30$. There is no $\hat{q} = 2$ phase because T_1 and T_2 are degenerate. Bottom: Group 1 has higher in-connectivity than the two others: $c_{11} = 30$, whereas $c_{22} = c_{33} = 15$. T_1 and T_2 are distinct, and from the recall scores we see that only group 1 is detected between T_1 and T_2, whereas groups 2 and 3 have an equally low recall score, as in the partition given by the algorithm, they are merged to a single group. Below T_2, $\hat{q} = 3$ and the algorithm separates groups 2 and 3. The spin-glass phase is not reached here.

merge some of the real groups together to obtain q groups. The more interesting case is when q is bigger than q^*.

First, the range of temperatures of the spin-glass phase grows as q increases. If ϵ is only slightly above the detectability threshold ϵ^* [24,26], then increasing q can lead to a situation where there is no recovery phase between the paramagnetic phase and the spin-glass phase.

However, we will focus on the case when ϵ is small enough for intermediate phases to be present. As described previously, the phase transitions are degenerate if p_{in} is the same for all groups. Therefore, we generally observe only one intermediate phase, with $\hat{q} = q^*$. However, this is not always the case and mod-bp can return partitions with different $\hat{q}$ values, depending on the intialization, similarily to what is observed on the real network in Fig. 3. Two phenomena can be observed, separately or simultaneously.

The first phenomenon is the one with $\hat{q} = q^* + 1$, where q^* of the groups correspond very well to the real groups, and the last group contains only a very small fraction of nodes. Depending on the initialization, this last group can even contain no node at all, in which case it can be simply discarded. This phenomenon is likely to come from the stochasticity of the SBM and is observed also for large networks with 10^5 nodes. The modularity of those partitions is equal to, or slightly higher than, those found in the $\hat{q} = q^*$ phase of mod-bp run with $q = q^*$, which explains why they are found. On the other hand, we have never observed more than one of these additional, and almost empty, groups, such that $\hat{q}$ is always at most equal to $q^* + 1$.

The other phenomenon is that of distinct groups merging together in the retrieval partition, leading to $\hat{q} < q^*$. Such partitions have lower modularities than partitions with $\hat{q} = q^*$ (found at same temperature from a different initialization), showing that the algorithm is unable to correctly maximize the modularity starting from any initialization. This is likely due to the existence of "hard but detectable" phases [24], in which frozen variables cause algorithms to be stuck in suboptimal solutions. A simple way out of this problem is to run the algorithm several times with different initial conditions, selecting, finally, the configuration of higher modularity found.

These two effects might coexist and produce retrieval partitions in which two of the real groups are merged into a single one and an additional group containing very few or even no nodes at all is also present. In this case $\hat{q} = q^*$, but the retrieval partition is not the right one. So the existence of an almost-empty group should be considered as a warning on the reliability of the mod-bp result.

C. Results on real networks

For community detection on real networks, q^* is in general unknown and there is no available ground truth. From Fig. 4 and the previous section, we know that mod-bp can converge to partitions with different $\hat{q}$ at the same temperature, depending on the initialization. This motivates us to run mod-bp several times for each temperature, which allows us to quantify the probability a given $\hat{q}$ is found at any given temperature T. Figure 5 shows the coexistence of phases in the "political books" [25] and "political blogs" [27] data sets for different values of q. The analysis made in these figures is similar to the one proposed in Ref. [28] for multiresolution community detection.

These figures suggest that, at a given normalized temperature T/T_0, the results returned by mod-bp only marginally depend on the chosen q as long as $q > q^*$. Moreover, we observe that, within a phase with a given number $\hat{q}$ of groups found, the partition $\{t\}$ only marginally depends on the temperature T. Averaging over the several partitions found at different temperatures and with different initial condition, we show in Fig. 6 (for "political books") and Fig. 7 (for "political blogs") that Q^{MAP} depends essentially on $\hat{q}$, and only minimally on q. As in ZM, we consider that the largest $\hat{q}$ leading to a significant increase of Q^{MAP} with respect to $\hat{q} - 1$ is a plausible estimate of q^*, which agrees well with the commonly accepted "ground truths" of $q^* = 3$ for "political books" and $q^* = 2$ for "political blogs."

In Fig. 7 we also show the distribution of overlaps between randomly chosen partitions with the same $\hat{q}$ for the "political blogs" data set. The overlap between two partitions $\{t\}$ and $\{s\}$ is a number between zero and 1 and is defined as

$$O(\{t\},\{s\}) = \max_{\sigma}\left[\frac{1}{N}\sum_{i=1}^{N}\delta_{t_i,\sigma(s_i)}\right], \tag{16}$$

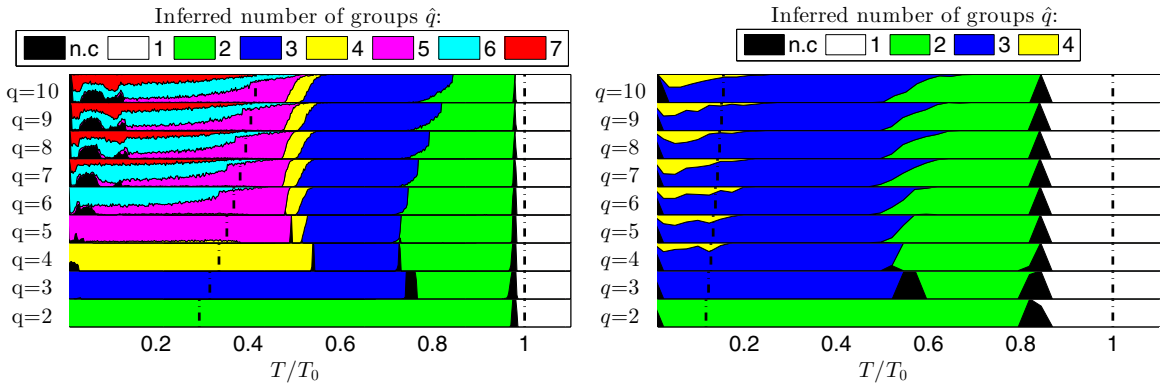

FIG. 5. (Color online) These plots show the inferred number of groups $\hat{q}$ as a function of the normalized temperature T/T_0 and of q for the "political books" (left plot) and "political blogs" (right plot) networks. The dotted lines mark $T = T^*$ (left line in each plot) and $T = T_0$ (right line in each plot). The "n.c" areas correspond to instances that did not reach the convergence criterion (10^{-6}) in 700 and 300 iterations, respectively, for the two networks. To take into account coexisting phases, the algorithm was run for 200 (respectively, 50) different initializations at each temperature. The position of T_1 is very stable across the different values of q and is characterized by a diverging number of iterations. The other critical temperatures T_k are not always well defined due to overlaps between phases and to phase transitions becoming much less sharp; however, up to $q = 4$, the phases stay well separated, with a clear divergence of the number of iterations at the phase boundaries. Remarkably, the existence domains of each phase in terms of T/T_0 does not vary a lot with q.

where the maximum over all permutations σ of $\{1,\ldots,\hat{q}\}$ allows us to lift the permutation symmetry of the group numbering choice. The closer the overlap between two partitions is to 1, the more similar they are. Figure 7 thus shows that partitions with the same $\hat{q}$ are very similar to one another, independently of the two parameters of mod-bp, T and q, for which they were obtained. One may be worried about the double peak structure of the $\hat{q} = 3$ case and wondering whether the two peaks do actually correspond to different communities structures. We have looked at the group partitions returned by the algorithm and found the following. There is always a well-conserved group of 520 to 530 nodes, while the remaining roughly 700 nodes can be clustered in different ways: For $\hat{q} = 3$, there are two different partitions with roughly 600+100 and 500+200 nodes; for $\hat{q} = 4$, the partition is roughly 380+280+40 nodes. All these configurations have essentially the same modularity. So the conclusion is that the $\hat{q} = 2$ partition (520+700 nodes) is significant, while further splitting of the cluster of 700 nodes is not very meaningful.

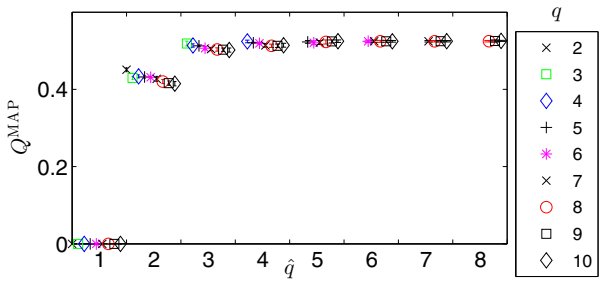

FIG. 6. (Color online) Q^{MAP} as a function of q and $\hat{q}$ for "political books" using the same experimental results as in Fig. 5. Symbols represent the mean Q^{MAP} of all experiments with a given q resulting in a given $\hat{q}$, along with an error bar representing the standard deviation. Despite the use of different temperatures, the error standard deviations are very small for each q. Furthermore, the mean Q^{MAP} for different q are very similar, such that we can consider Q^{MAP} to essentially depend on $\hat{q}$, with only negligible influence of q and T. The fact that the increase in Q^{MAP} for $\hat{q} > 3$ is minimal agrees with the fact that $q^* = 3$.

D. Results on hierarchical networks

To validate our results on a hierarchical network, we ran mod-bp on the "air transportation network," which is a network of cities in which an edge is present between each pair of

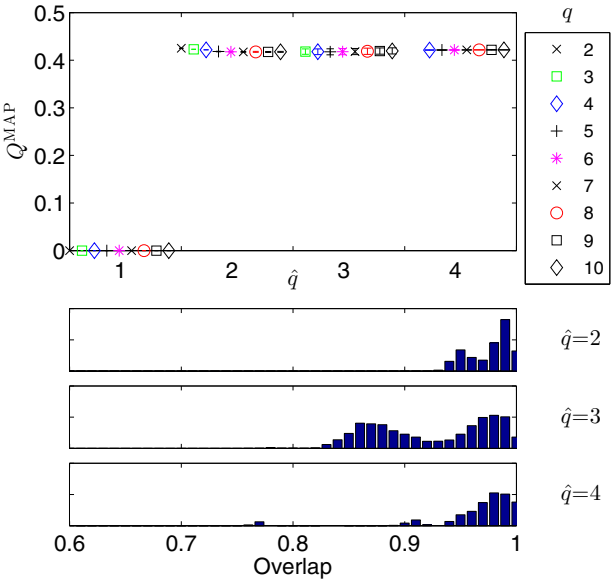

FIG. 7. (Color online) Top: Q^{MAP} as a function of q and $\hat{q}$ for "political blogs" using the same experimental results as in Fig. 5. Symbols represent the mean Q^{MAP} of all experiments with a given q resulting in a given $\hat{q}$, along with an error bar representing the standard deviation. The fact that Q^{MAP} does not increase for $\hat{q} > 2$ agrees with the fact that $q^* = 2$. Bottom: Empirical distribution of 20 000 overlaps between pairs of partitions with same $\hat{q}$.

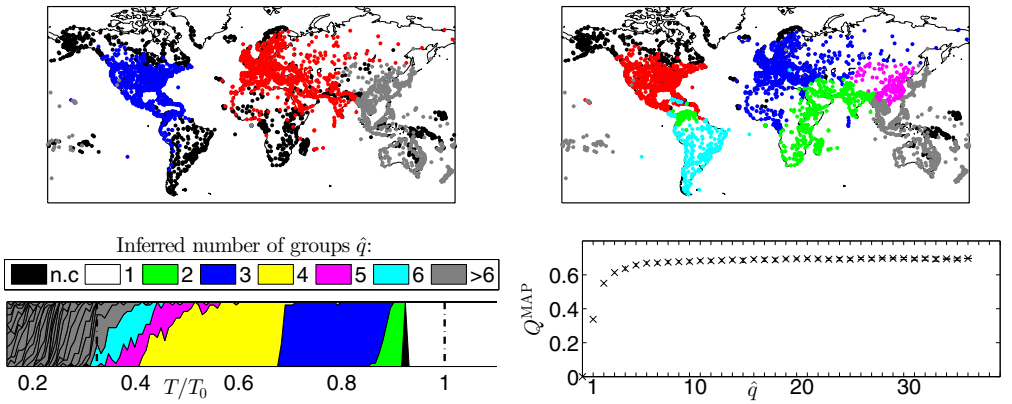

FIG. 8. (Color online) Clustering of cities in the "air transportation network" [29] using mod-bp with $q = 50$. Top left: Clustering into four communities ($T/T_0 = 0.6$). Top right: Clustering into seven communities ($T/T_0 = 0.35$). The communities mainly correspond to geographical and geopolitical units. Increasing the number of communities, substructures of bigger entities appear (e.g., China separates from the east-asian cluster). Lower-degree nodes are initially placed in the same group (e.g., Alaska and Madagascar are initially in the same cluster), but lowering T lets new clusters appear (e.g., South America). Bottom left: As in Fig. 5, we show the frequency of retrieval configurations with different $\hat{q}$ as a function of T/T_0 (on the base of 20 runs per temperature). While phases up to $\hat{q} \approx 7$ exist in broad ranges of temperatures, phases with higher $\hat{q}$ exist on much narrower ranges and coexist with many different other phases, which makes it unclear which $\hat{q}$ is more meaningful than others. Bottom right: Q increases only marginally with $\hat{q}$ for $\hat{q} \gtrsim 7$.

cities connected by direct flights [29,30]. A coarse-grained clustering results in a few communities of cities that are well connected to each other. Each of these communities corresponds to geographical and geopolitical units that are clearly recognizable, which can be further subdivided into subcommunities. For example, the US and Mexico are two subcommunities of the "North America" cluster. We ran mod-bp with $q = 50$ for temperatures from 0 to $1.2 \times T_0$ and present the results in Fig. 8. As expected, the number of found communities increases with decreasing T/T_0, thus revealing substructures with increasing geographical precision. Based on the modularity and the temperature range of the phases, $\hat{q} \approx 7$ seems to be a meaningful number of communities. Further decreasing the temperature splits the communities into smaller ones, and individual countries appear as single or even several communities.

IV. DISCUSSION

In addition to not requiring the knowledge of the generative model, a further advantage of mod-bp is that it has only two adjustable parameters, T and q. However, for a given network, it is not clear how to choose them in order to obtain the optimal partition. The recommendation of ZM is to run mod-bp at $T^*(q)$, defined in Eq. (7), for increasing values of q, until it no longer leads to a significant increase in modularity. Based on the experiments on synthetic and real networks presented in this work, we conclude that an important additional step in this procedure is to calculate the effective number of groups $\hat{q}$ of each partition returned by the algorithm, which can differ from q. Furthermore, this phenomenon leads to a new rule for assigning a group to each node, given that some groups might be merged, which also affects the modularity.

Another possible way to proceed is to run mod-bp with a large value of q and sweep the temperature scale from $T_0(q)$ downwards. As T is lowered, the network is clustered into an increasing number of effective groups $\hat{q}$ and the found partitions have increasing modularities. Again, the procedure can be stopped once the modularity no longer increases in a significant way as $\hat{q}$ is increased.

For real networks, where the generating process is in general not known and not as straightforward as in the SBM, the number of groups to cluster the nodes is in part left as a choice to the user. In this case, running mod-bp with a quite large value of q and using T as the parameter to search for the optimal partition seem both desirable and efficient. To make the optimal choice, in addition to the value of the modularity of a partition with $\hat{q}$ groups, the range of temperatures where this $\hat{q}$ phase exists might indicate how relevant it is (as shown in Fig. 5). In particular, if a $\hat{q}$ phase only exists on a narrow range of temperatures, then it is likely to be less important, because it is less stable with respect to changes in the model parameter (T in the present case).

Furthermore, as seen on graphs generated by the SBM, it may be that some groups contain a very small number of nodes. In this case, merging them with bigger groups will only slightly change the modularity and result into a more meaningful and stable partition.

V. CONCLUSION

In this paper, we have studied the mod-bp algorithm proposed in Ref. [11], focusing on the influence of the choice of the two adjustable parameters q and T on both real and synthetic networks. We have given a more precise picture of the algorithm behavior by identifying new order parameters that allow us to define several different phases inside the recovery phase. In each of these phases, mod-bp clusters the nodes into a different number of groups $\hat{q}$. These phases can either be well separated on the temperature scale and be accompanied by a divergence in the number of iterations of the algorithm or coexist in the low-temperature regime. The partitions with the same number $\hat{q}$ of groups typically have high overlaps

among them and very similar modularities. We have proposed a normalized temperature scale (T/T_0) on which mod-bp has a very similar behavior for different values of q. These findings provide a broader description of the mod-bp algorithm behavior, showing its robustness and effectiveness. Hopefully they can be very useful when mod-bp is run on real networks where the ground truth is unknown.

Real networks may have hierarchical structures [28,31–33] and the deeper understanding of the different recovery phases achieved in this work may help in using the temperature as a simple parameter to study by mod-bp different levels of the hierarchical clustering. The different levels of clustering hierarchy may correspond to recovery phases with different values of $\hat{q}$, obtained decreasing the temperature.

ACKNOWLEDGMENTS

The authors thank Cristopher Moore, Lenka Zdeborova, and Pan Zhang for useful discussions. This research has received funding from the Italian Research Ministry through FIRB Project No. RBFR086NN1. C.S. was funded by the Université franco-italienne.

[1] D. Hric, R. K. Darst, and S. Fortunato, Phys. Rev. E **90**, 062805 (2014).
[2] P. W. Holland, K. B. Laskey, and S. Leinhardt, Soc. Networks **5**, 109 (1983).
[3] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, Phys. Rev. Lett. **107**, 065701 (2011).
[4] M. B. Hastings, Phys. Rev. E **74**, 035102 (2006).
[5] M. E. J. Newman, Phys. Rev. E **74**, 036104 (2006).
[6] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang, Proc. Natl. Acad. Sci. USA **110**, 20935 (2013).
[7] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).
[8] J. Duch and A. Arenas, Phys. Rev. E **72**, 027104 (2005).
[9] D. Aloise, S. Cafieri, G. Caporossi, P. Hansen, S. Perron, and L. Liberti, Phys. Rev. E **82**, 046112 (2010).
[10] S. Cafieri, P. Hansen, and L. Liberti, Phys. Rev. E **83**, 056105 (2011).
[11] P. Zhang and C. Moore, Proc. Natl. Acad. Sci. USA **111**, 18144 (2014).
[12] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[13] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner, IEEE Transact. Knowl. Data Eng. **20**, 172 (2008).
[14] P. Erdős and A. Rényi, Bull. Inst. Internat. Statist. **38**, 343 (1961).
[15] R. Guimera, M. Sales-Pardo, and L. A. N. Amaral, Phys. Rev. E **70**, 025101 (2004).
[16] J. Reichardt and S. Bornholdt, Phys. Rev. E **74**, 016110 (2006).
[17] A. Lancichinetti, F. Radicchi, and J. J. Ramasco, Phys. Rev. E **81**, 046110 (2010).
[18] S. Fortunato and M. Barthélemy, Proc. Natl. Acad. Sci. USA **104**, 36 (2007).
[19] M. Mézard and A. Montanari, *Information, Physics, and Computation* (Oxford Press, Oxford, 2009).
[20] J. S. Yedidia, W. T. Freeman, and Y. Weiss, in *Exploring Artificial Intelligence in the New Millennium* (Morgan Kaufmann Publishers Inc., 2003), pp. 239–269.
[21] A. Lancichinetti and S. Fortunato, Sci. Rep. **2**, 336 (2012).
[22] D. Hu, P. Ronhovde, and Z. Nussinov, Philos. Mag. **92**, 406 (2012).
[23] P. Ronhovde, D. Hu, and Z. Nussinov, Europhys. Lett. **99**, 38006 (2012).
[24] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, Phys. Rev. E **84**, 066106 (2011).
[25] "Books about US politics", http://networkdata.ics.uci.edu/data.php?d=polbooks.
[26] E. Mossel, J. Neeman, and A. Sly, arXiv:1202.1499 (2012).
[27] L. A. Adamic and N. Glance, in *Proceedings of the 3rd International Workshop on Link Discovery* (ACM, New York, 2005), pp. 36–43.
[28] P. Ronhovde and Z. Nussinov, Phys. Rev. E **80**, 016109 (2009).
[29] R. Guimera, S. Mossa, A. Turtschi, and L. N. Amaral, Proc. Natl. Acad. Sci. USA **102**, 7794 (2005).
[30] M. Sales-Pardo, R. Guimera, A. A. Moreira, and L. A. N. Amaral, Proc. Natl. Acad. Sci. USA **104**, 15224 (2007).
[31] M. Girvan and M. E. Newman, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002).
[32] T. P. Peixoto, Phys. Rev. X **4**, 011047 (2014).
[33] Z. Nussinov, P. Ronhovde, D. Hu, S. Chakrabarty, M. Sahu, B. Sun, N. Mauro, and K. Sahu, arXiv:1503.01626 (2015).