Inversion method for content-based networks

José J. Ramasco*
CNLL, ISI Foundation, Viale S. Severo 65, I-10133 Torino, Italy

Muhittin Mungan†
Department of Physics, Faculty of Arts and Sciences, Boğaziçi University, 34342 Bebek, Istanbul, Turkey and The Feza Gürsey Institute, P.O.B. 6, Çengelköy, 34680 Istanbul, Turkey

(Received 14 November 2007; published 27 March 2008)

In this paper, we generalize a recently introduced expectation maximization (EM) method for graphs and apply it to content-based networks. The EM method provides a classification of the nodes of a graph, and allows one to infer relations between the different classes. Content-based networks are ideal models for graphs displaying any kind of community and/or multipartite structure. We show both numerically and analytically that the generalized EM method is able to recover the process that led to the generation of such networks. We also investigate the conditions under which our generalized EM method can recover the underlying content-based structure in the presence of randomness in the connections. Two entropies, S_q and S_c, are defined to measure the quality of the node classification and to what extent the connectivity of a given network is content based. S_q and S_c are also useful in determining the number of classes for which the classification is optimal.

DOI: 10.1103/PhysRevE.77.036122 PACS number(s): 89.75.Hc, 02.50.Tt

I. INTRODUCTION

Classifying items with respect to their properties is a fundamental and very old problem. If the properties are inherent to the objects, the difficulty is deciding first how many groups are required and then establishing the discrimination thresholds for each. The matter becomes more complicated when instead of the inherent properties, one tries to classify elements based on mutual interactions. Of course, such classifications would be very useful for a better understanding of the mechanisms underlying the behavior of systems encountered in scientific disciplines as diverse as sociology, biology or physics [1–4]. As an example, consider social systems which are often modeled as networks. The vertices represent individuals and the edges interactions between them. These interactions can be of many types: friendship, belonging to the same club or school, working together, etc. In these graphs, it is important to be able to group the nodes into what is commonly known as *communities*. That is, groups of vertices that share a higher number of connections among themselves than with the rest of the network [5–9] (see also [10] for a recent review). This partition bears information on which persons have a stronger interdependence and may allow one to predict the actors that drive the dynamics of the group as a whole. In biology, on the other hand, network methods have been used to understand gene regulatory patterns [11]. Here, each vertex corresponds to a gene and an edge contains information on how the associated protein regulates the synthesis of the protein associated to the second gene. Since regulation of gene activity plays a fundamental role in the functioning of the cell [12], the community structure points towards the different functional subunits (see [13] and references therein). Given the relevance of communities, recent years have seen an increase in the number of techniques proposed to detect them. To name a few: some of them are based on the concept of betweenness (number of paths passing through a link) and modularity [8,9,14], others on synchronization of oscillators [15,16] or on other dynamical systems running on the network [17–19], detection of overlapping cliques [20] or the diffusion of random walkers [21–23].

Nevertheless, communities are not the only relevant information that can be extracted from networks. It is also possible to search for vertices with similar connection patterns (not necessarily having connections among themselves, as in the case of communities) that are expected to play equivalent functional roles. In the social networks literature, such nodes are referred to as *structurally equivalent* [24] and have led to an analysis of social networks based on *block modeling* [1,25]. In many types of networks, such as those formed by webpages or social actors, the connection between nodes is often due to some intrinsic properties of the nodes, which we will refer to henceforth as their "contents." Thus it is possible to consider an alternative point of view in which a network structure arises as a result of node contents, leading to the notion of content-based networks [26–29].

In many cases, network analysis approaches based on communities and those based on some form of node similarity are aimed towards the understanding of very different questions. When viewed within the framework of content-based networks, however, these differences disappear, as will be argued below. We will also show that an extension of Newman and Leicht's expectation maximization (EM) method [30] is well suited for uncovering content-based structure underlying a network, inverting in practice the process that led to its formation. We will define as well two entropies, Sq and S_c, that are useful in measuring the quality of an EM classification. These entropies provide a way of determining the number of classes for which the classification is optimal.

*jramasco@isi.it
†mmungan@boun.edu.tr

The organization of the paper is as follows: in Sec. II, content-based networks are formally introduced. Next, we describe in Sec. III our generalization of the EM method to directed graphs. In Sec. IV, we show how the EM method can be used to solve the inverse problem, namely to recover the underlying content-based structure from a given network. We present in Sec. V analytical results regarding the application of the EM method to content-based networks and the recovery of the content-based structure. These results will be complemented with a numerical study in Secs. VI and VII. In Sec. VII, we consider a more realistic situation and ask to what extent an underlying content-based structure can be recovered in the presence of disorder in the connections. Finally, we summarize our results and present the conclusions in Sec. VIII.

II. CONTENT-BASED NETWORKS

Let us define first content-based networks. Consider a set of nodes $i=1,2,\ldots N$, each of which has a *content* x_i assigned with $x_i \in \mathcal{X}=\{1,2,\ldots,\mathcal{N}_x\}$, and where $1,2,\ldots$ are labels for the possible contents. The structure of the connectivity pattern of the associated content-based network is determined by the function $c(x_i,x_j) \in \{0,1\}$, which is defined for all ordered pairs of contents $(x,y) \in \mathcal{X}$. The adjacency matrix of the graph is then given by

$$A_{ij} = c(x_i,x_j). \tag{1}$$

We see immediately that nodes having the same contents x also have the same connection patterns, and thus are structurally equivalent [24]. As explained before, this can imply a functional equivalence in the process that generated the network. The point of view that we will take in this article is to regard content-based networks as ideal networks, from which the "real" networks are obtained through alteration or removal of some of the connections. Note that the range of topologies that can be generated via content-based network is very broad: if the connectivity function $c(x,y)$ shows a close to diagonal configuration, the network will be formed by a set of almost insulated cliques. The ideal configuration would be a family of independent communities without interconnections. Another configuration that can be easily reproduced with content-based networks are bipartite graphs. In its most simplest form, it is enough to allow the nodes to take one of two possible contents and let the connectivity function c to be nonzero only for the off-diagonal elements. Much more complicated connectivity patterns can be actually achieved by introducing finer contents distinctions and more intricate connectivity functions. Thus a content-based graph can in general include all sorts of combinations between communitylike and/or multipartite graphs, as can be seen in the example plotted in Fig. 1.

Another point to note is that originally these networks were proposed in a context where the relation between contents was an order relation [26,31,32]. This implies that the relation between nodes is not symmetric and the network is therefore more naturally represented by a directed graph. In this case, the connectivity function c is nonsymmetric in its arguments. Apart from directionality, realistic graphs may present, as well, a certain degree of disorder in their connection patterns. This effect can be incorporated into the mathematical description by regarding the values of $c(x,y)$ as probabilities of having a link from a node of content x to a node of content y. This view transforms the content-based network into a hidden variable graph [33–35]. As we will see later, the EM method is still able to extract content information from networks produced in this way but the failure rate increases the further $c(x,y)$ deviates from a binary-valued function.

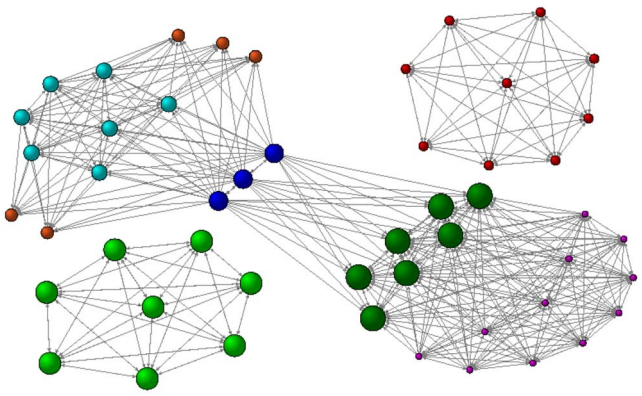

FIG. 1. (Color online) An example of a content-based network, the colors and the sizes of the nodes correspond to the different contents (green A, red B, blue C, magenta D, cyan E, olive F, and orange G).

Contents based networks have proven to be very useful in the description of phenomena that include an underlying relation of hierarchy or ordering. The simplest way of achieving such a relation is to associate with each node a string of letters and letting the relation between any two nodes be based on string inclusion: namely that one string is contained as an uninterrupted subsequence in the other. Networks generated from random strings in this manner have been successfully used to model receptor-ligand interactions in the immune system [31,32], and the transcription factor based gene regulatory network in yeast [26–29].

In this paper, our goal is to address the *inverse* problem: Given a network of which we know nothing in advance, is it possible to decide whether there is an underlying content-based structure and, if so, can we deduce the class membership of its nodes and the class connectivity function? Moreover, can this be achieved in the presence of noisy connections? Seen in this way, the problem at hand becomes one of statistical inference, very well suited to EM methods [36,37].

III. THE EM METHOD FOR NETWORKS AND ITS GENERALIZATION

Given a graph $\mathcal{G}$ of N nodes and an adjacency matrix A_{ij}, the expectation maximization (EM) algorithm [30] searches for a partition of the nodes into $\mathcal{N}_c$ groups such that a certain log-likelihood function for the graph is maximized. Henceforth we will refer to the groups into which the EM method divides the nodes, as *classes*. Note that $\mathcal{N}_c$ must not be con-

fused with the number of contents $\mathcal{N}_x$, described in the previous section. Ideally, the optimal number of classes would be $\mathcal{N}_x$, but a criterion independent from the EM algorithm is required to determine first its value, since in general $\mathcal{N}_x$ will not be known in advance. We will offer such a criterion in the next section. The variables of the EM algorithm are the probabilities π_r that a randomly selected node is assigned to class r, with $r=1,2,\ldots\mathcal{N}_c$, and the set of probabilities θ_{rj} of having a connection from a node belonging to class r to a certain node j. Assuming that the functions θ and π are given, the probability $\Pr(A,g|\pi,\theta)$ of realizing the given graph under a node classification g, such that g_i is the class that node i has been assigned to, can be written as

$$\Pr(A,g|\pi,\theta)=\prod_i \pi_{g_i}\left[\prod_j \theta_{g_i,j}^{A_{ij}}\right]. \tag{2}$$

$\Pr(A,g|\pi,\theta)$ is the likelihood to be maximized, but it turns out to be more convenient to consider its logarithm instead:

$$\mathcal{L}(\pi,\theta)=\sum_i \left[\ln \pi_{g_i}+\sum_j A_{ij}\ln \theta_{g_i,j}\right]. \tag{3}$$

Treating the *a priori* unknown class assignment g_i of the nodes as statistical "unknown data," one introduces next the auxiliary probabilities $q_{ir}=\Pr(g_i|A,\pi,\theta)$ that a node i is assigned to class r, and considers the averaged log-likelihood constructed as

$$\bar{\mathcal{L}}(\pi,\theta)=\sum_{ir} q_{ir}\left[\ln \pi_r+\sum_j A_{ij}\ln \theta_{rj}\right]. \tag{4}$$

The maximization of $\bar{\mathcal{L}}$ must be performed taking into account the following normalization conditions for the probabilities π and θ

$$\sum_{r=1}^{\mathcal{N}_c}\pi_r=1, \tag{5}$$

$$\sum_{j=1}^{N}\theta_{rj}=1. \tag{6}$$

The final results are

$$\pi_r=\frac{1}{N}\sum_i q_{ir}, \tag{7}$$

$$\theta_{rj}=\frac{\sum_i A_{ij}q_{ir}}{\sum_i k_i q_{ir}}, \tag{8}$$

where k_i is the out-degree of node i. The still unknown probabilities q_{ir} are then determined *a posteriori* by noting that

$$q_{ir}=\Pr(g_i=r|A,\pi,\theta)=\frac{\Pr(A,g_i=r|\pi,\theta)}{\Pr(A|\pi,\theta)}, \tag{9}$$

from which one obtains

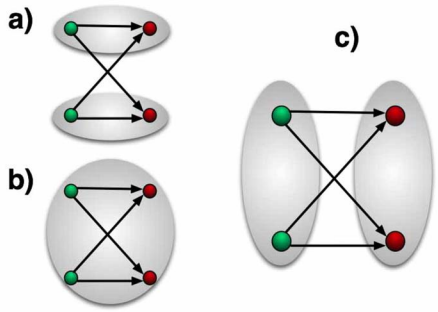

FIG. 2. (Color online) A simple scenario in which the EM method for directed networks, as defined in [30], has problems in classifying the nodes of the network in two classes. The configurations (a) and (b) are possible outputs of the original EM method since both satisfy the normalization condition of Eq. (6). The solution (a) comes together with values for $q_{ir}=1/2$ for all the nodes and classes, while the solution (b), which has a lower likelihood, produces $q_{ir}\approx 0.99$ for all the nodes in one class and a very small probability for the other. The plot on the right, solution (c), is the output offered by the generalization of EM with values of q_{ir} virtually one or zero.

$$q_{ir}=\frac{\pi_r\prod_j \theta_{rj}^{A_{ij}}}{\sum_s \pi_s\prod_j \theta_{sj}^{A_{ij}}}. \tag{10}$$

Eqs. (7), (8), and (10) form a set of self-consistent equations for q_{ir}, θ_{rj}, and π_r that any extremum of the expected log-likelihood must satisfy.

Thus, given a graph $\mathcal{G}$, the EM algorithm consists of picking a number of classes $\mathcal{N}_c$ into which the nodes are to be classified and searching for solutions of Eqs. (7), (8), and (10). These equations were derived by Newman and Leicht [30]. They also showed that when applied to diverse type of networks, the resulting q_{ir} and θ_{rj} yield useful information about the internal structure of the network. Note that only a minimal amount of a priori information is supplied: the number of classes $\mathcal{N}_c$ and the network.

However, the EM method in the form presented so far does not yet serve our purposes for the following reason: as noted previously, content-based networks are usually represented as directed graphs. The probability θ_{rj} was defined as the probability that a node j receives a directed connection from a node belonging to class r. Together with the normalization condition for θ_{rj}, Eq. (6), this implies that the classification must be such that each class r has at least one member with nonzero out-degree. This constraint forces the EM algorithm to classify a simple bipartite graph in the manner depicted in Figs. 2(a) or 2(b). From a content-based point of view, on the other hand, the classification that would be more natural is the one displayed in Fig. 2(c) which is forbidden by the condition of Eq. (6). This difficulty is not resolved by redefining θ_{rj} instead, as the probability that a node j makes a directed connection *to* a node belonging to class r, since now the classification must be such that each class r has at least one member with nonzero in-degree.

We therefore have to generalize the EM approach in such a way that the node directionality does not restrict the possible classification of the nodes. This can be achieved by introducing the following probabilities.

(i) $\vec{\theta}_{ri}$ of having a unidirectional link from a vertex of class r to a node i;

(ii) $\overleftarrow{\theta}_{ri}$ of having a unidirectional link from node i to a node in class r; and

(iii) $\overleftrightarrow{\theta}_{ri}$ of having a bidirectional link between i and a node in class r.

With these new definitions, Eq. (2) becomes

$$\Pr(A,g|\pi,\overleftarrow{\theta},\vec{\theta},\overleftrightarrow{\theta}) = \prod_i \left[\pi_{g_i} \prod_j \overleftarrow{\theta}_{g_i,j}^{A_{ji}(1-A_{ij})} \vec{\theta}_{g_i,j}^{A_{ij}(1-A_{ji})} \overleftrightarrow{\theta}_{g_i,j}^{A_{ij}A_{ji}} \right]. \tag{11}$$

The likelihood can now be written as

$$\bar{\mathcal{L}}(\pi,\theta) = \sum_{ir} q_{ir} \Bigg(\ln \pi_r + \sum_j [A_{ji}(1-A_{ij}) \ln \overleftarrow{\theta}_{r,j} + A_{ij}(1-A_{ji}) \ln \vec{\theta}_{r,j} + A_{ij}A_{ji} \ln \overleftrightarrow{\theta}_{r,j}] \Bigg), \tag{12}$$

which has to be maximized under the following constraint on the probabilities θ_{rj}:

$$\sum_i (\overleftarrow{\theta}_{r,i} + \vec{\theta}_{r,i} + \overleftrightarrow{\theta}_{r,i}) = 1, \tag{13}$$

implying that there is no isolated node. The probability π_r, that a randomly selected node belongs to class r, is again given by Eq. (7).

Introducing the Lagrange multipliers β and λ_r, to incorporate the constraints, Eqs. (5) and (13), the expression to be extremized, becomes

$$\tilde{\bar{\mathcal{L}}} = \bar{\mathcal{L}} + \beta \Bigg(1 - \sum_r \pi_r \Bigg) + \sum_r \lambda_r \Bigg(1 - \sum_i (\overleftarrow{\theta}_{r,i} + \vec{\theta}_{r,i} + \overleftrightarrow{\theta}_{r,i}) \Bigg). \tag{14}$$

As before, the extremal condition on $\tilde{\bar{\mathcal{L}}}$ with respect to π gives us

$$\frac{\partial \tilde{\bar{\mathcal{L}}}}{\partial \pi_r} = 0 \Leftrightarrow \pi_r = \frac{1}{N} \sum_i q_{ir} \quad \text{and} \quad \beta = N, \tag{15}$$

where N is the total number of nodes. Differentiating $\tilde{\bar{\mathcal{L}}}$ with respect to the θ variables, we get [38]

$$\frac{\partial \tilde{\bar{\mathcal{L}}}}{\partial \overleftarrow{\theta}_{rj}} = 0 \Leftrightarrow \sum_i q_{ir} A_{ji}(1-A_{ij}) - \overleftarrow{\theta}_{rj} \lambda_r = 0,$$

$$\frac{\delta \tilde{\bar{\mathcal{L}}}}{\delta \vec{\theta}_{rj}} = 0 \Leftrightarrow \sum_i q_{ir} A_{ij}(1-A_{ji}) - \vec{\theta}_{rj} \lambda_r = 0,$$

$$\frac{\delta \tilde{\bar{\mathcal{L}}}}{\delta \overleftrightarrow{\theta}_{rj}} = 0 \Leftrightarrow \sum_i q_{ir} A_{ij} A_{ji} - \overleftrightarrow{\theta}_{rj} \lambda_r = 0. \tag{16}$$

Putting together the three previous expressions and summing over the index of the nodes j, we obtain the following result for the Lagrange multipliers:

$$\lambda_r = \sum_i q_{ir} (\bar{k}_i^i + \bar{k}_i^o - \bar{k}_i^b), \tag{17}$$

where $\bar{k}_i^i$, $\bar{k}_i^o$, and $\bar{k}_i^b$ are the in-degree, out-degree, and bidirectional degree of node i, respectively. Inserting this relation into the previous set of equations, we extract the new extremal conditions for the θ's:

$$\overleftarrow{\theta}_{rj} = \frac{\sum_i q_{ir} A_{ji}(1-A_{ij})}{\sum_i q_{ir}(\bar{k}_i^i + \bar{k}_i^o - \bar{k}_i^b)},$$

$$\vec{\theta}_{rj} = \frac{\sum_i q_{ir} A_{ij}(1-A_{ji})}{\sum_i q_{ir}(\bar{k}_i^i + \bar{k}_i^o - \bar{k}_i^b)},$$

$$\overleftrightarrow{\theta}_{rj} = \frac{\sum_i q_{ir} A_{ij} A_{ji}}{\sum_i q_{ir}(\bar{k}_i^i + \bar{k}_i^o - \bar{k}_i^b)}. \tag{18}$$

These expressions have to be again supplemented with the self-consistent equation for q_{ir} which now reads

$$q_{ir} = \frac{\pi_r \prod_j \overleftarrow{\theta}_{rj}^{A_{ji}(1-A_{ij})} \vec{\theta}_{rj}^{A_{ij}(1-A_{ji})} \overleftrightarrow{\theta}_{rj}^{A_{ij}A_{ji}}}{\sum_s \pi_s \prod_j \overleftarrow{\theta}_{sj}^{A_{ji}(1-A_{ij})} \vec{\theta}_{sj}^{A_{ij}(1-A_{ji})} \overleftrightarrow{\theta}_{sj}^{A_{ij}A_{ji}}}. \tag{19}$$

Note that when we have only bidirectional links so that $A_{ij}=A_{ji}$, it follows from Eq. (18) that $\overleftarrow{\theta}_{rj}=\vec{\theta}_{rj}=0$. Thus we recover the original EM equations under the identification $\theta_{rj}=\overleftrightarrow{\theta}_{rj}$.

It is easily shown that the solutions of the EM equations, Eqs. (7), (18), and (19), are such that if two nodes i and j are structurally equivalent, i.e. $A_{ik}=A_{jk}$ as well as $A_{ki}=A_{kj}$, for all k then they will be classified in the same manner: $q_{ir}=q_{jr}$, and $\overleftarrow{\theta}_{ri}=\overleftarrow{\theta}_{rj}$, $\vec{\theta}_{ri}=\vec{\theta}_{rj}$ and $\overleftrightarrow{\theta}_{ri}=\overleftrightarrow{\theta}_{rj}$ for all r. This property of the solutions obtained from the EM methods renders it very well suited for detecting any underlying content-based structure.

IV. THE INVERSION METHOD

One important shortcoming of the EM method is that $\mathcal{N}_c$ has to be provided as an external parameter. The algorithm lacks a means to evaluate how good a classification is, and consequently one cannot decide which number of classes furnishes an optimal classification of the nodes of a graph. To overcome this problem, we propose to define a measure of the quality of a classification as follows:

$$S_q \equiv -\frac{1}{N}\sum_{i,r} q_{ir} \ln q_{ir}, \tag{20}$$

where the sum runs over all the nodes i and classes r. S_q is the average entropy of the classification and as such measures the certainty with which the nodes are assigned to their respective classes. One can easily see that $0 \leq S_q \leq \ln \mathcal{N}_c$. For a sharp classification $S_q=0$, while the worst-case scenario occurs when $q_{ir}=1/\mathcal{N}_c$. We will later argue that S_q is a useful statistic to infer the optimal $\mathcal{N}_c$.

Once an optimal classification has been found, it is possible to determine the connectivity structure among the classes. Given an EM classification, we will define $\tilde{c}(r,s)$ as the probability that a node in class r has a connection to one in class s. This probability can be estimated as

$$\tilde{c}(r,s) = \frac{\sum_{ij} q_{ir} A_{ij} q_{js}}{\sum_i q_{ir} \sum_j q_{js}} \left(1 - \frac{\delta_{rs}}{1 - \sum_i q_{ir}}\right), \tag{21}$$

by noting that

$$p(i|r) = \frac{q_{ir}}{\sum_j q_{jr}} \tag{22}$$

is the posterior probability that given that a node has been assigned to class r, the node is i. The second term on the right-hand side of Eq. (21) must be included as a correction for the absence of self-connections, since by convention, we assume that $A_{ii}=0$ for all i.

$\tilde{c}(r,s)$, as defined above, is the probability of regarding a connection between two nodes in the graph as being one between nodes of type r and s. As we will show in the following section, if the underlying graph is a content-based network, a successful application of the EM algorithm should result in sharp assignments of nodes into classes and $\tilde{c}(r,s)$ should thus be binary valued [and moreover be equal to the connectivity function $c(r,s)$]. It is possible to also define a measure of how close the connectivity function resembles one that corresponds to a content-based network by considering the entropy for $\tilde{c}$,

$$S_c \equiv -\frac{2}{\mathcal{N}_c^2 \ln 2}\sum_{rs} \tilde{c}(r,s) \ln \tilde{c}(r,s). \tag{23}$$

We have that $0 \leq S_c \leq 1$. The maximum of S_c occurs when $\tilde{c}(r,s)=1/2$, i.e. when none of the classes have any preferred connection pattern to any class.

The generalization of the EM method, the entropies S_q, S_c, and the estimation of $\tilde{c}(r,s)$ are in general applicable to any kind of graph. However, for the purpose of this paper we will focus only on their applications to content-based networks. We will address the general case in a subsequent work [39], where we will also show that content-based networks play a special role for the classifications of the EM method.

V. ANALYTICAL RESULTS FOR CONTENT-BASED NETWORKS

Assume that we are given a content-based graph $\mathcal{G}$ that has been constructed from a set of nodes of unknown contents, and an unknown connectivity function $c(x,y)$. In this setting, we suppose that the optimal number of classes $\mathcal{N}_c$ has already been found and that it is equal to the number of contents $\mathcal{N}_x$. We would like to know under which conditions the EM algorithm can infer the class membership of the nodes as well as the connectivity function. In other words, given the adjacency matrix A_{ij}, we are looking for a solution of the generalized EM equations, Eqs. (18) and (19), with

$$q_{ir} = \delta_{r,x_i} \quad \text{with} \quad x_i \in \mathcal{X}, \tag{24}$$

along with the unknown class-connectivity function $\tilde{c}(r,s)$ that ideally should coincide with the original $c(x,y)$. Note that the ansatz Eq. (24) implies that for such a solution $S_q=0$.

Substituting the above ansatz into Eq. (18), we find

$$\overleftarrow{\theta}_{rj} = \frac{c(x_j,r)[1-c(r,x_j)]}{\bar{k}_r^i + \bar{k}_r^o - \bar{k}_r^b},$$

$$\vec{\theta}_{rj} = \frac{c(r,x_j)[1-c(x_j,r)]}{\bar{k}_r^i + \bar{k}_r^o - \bar{k}_r^b},$$

$$\overleftrightarrow{\theta}_{rj} = \frac{c(r,x_j)c(x_j,r)}{\bar{k}_r^i + \bar{k}_r^o - \bar{k}_r^b}, \tag{25}$$

where $\bar{k}_r^i$, $\bar{k}_r^o$, and $\bar{k}_r^b$ are the average in-degree, out-degree and bidirectional degree of nodes belonging to class r,

$$\sum_i \delta_{x_i,r}(\bar{k}_i^i + \bar{k}_i^o - \bar{k}_i^b) = n_r(\bar{k}_r^i + \bar{k}_r^o - \bar{k}_r^b) \equiv n_r \bar{k}_r, \tag{26}$$

so that $\bar{k}_r$ is the total degree of each of the n_r nodes belonging to class r. Note that in Eq. (25), the node index j enters only through its content x_j, so that θ_{rj} is the same for all the nodes that have the same content as j. The same turns out to be true for the q_{ir}. We thus have $q_{ir}=q_{tr}$ for all nodes i such that $x_i=t$, and from Eq. (19) we obtain

$$\begin{aligned} q_{tr} = \frac{\gamma_t \pi_r}{\bar{k}_r^{\bar{k}_t}} \prod_s \{&[c(r,s)(1-c(s,r))]^{c(t,s)(1-c(s,t))} \\ &\times [c(s,r)(1-c(r,s))]^{c(s,t)(1-c(t,s))} \\ &\times [c(r,s)c(s,r)]^{c(t,s)c(s,t)}\}, \end{aligned} \tag{27}$$

where γ_t is the normalization constant for q_{tr}.

We now have to consider the conditions on $c(r,s)$, $c(s,r)$, $c(t,s)$, and $c(s,t)$ such that given the classes r and t, the terms in the product on the right-hand side of Eq. (27) are nonzero for all s, when $r=t$, and zero for at least one s when $r \neq t$. This is a statement about the kind of connections that the nodes of type r and t make to or receive from nodes of all possible classes s. An inspection of the c^c-type

terms in the product shows that their contribution to q_{tr} is nonzero if and only if the following two conditions are satisfied for all s:

(i) If there is a connection between t and s, there must be also a connection between r and s of the same kind, namely either in, out, or bidirectional.

(ii) Whenever there is no connection between t and s, there can be any kind of connection between r and s, as well as none at all.

The satisfaction of both conditions can be regarded as constituting a cover type of relation between r and t, i.e. nodes belonging to class r connect in the same way with all the classes that nodes belonging to class t connect, *but* they have also some extra connections. We will denote this relation by $r \succ t$ and say that r covers t. From its definition it is clear that the cover relation is transitive, $r \succ t, t \succ s \Rightarrow r \succ s$. When $r \succ t$, we also define $\mathcal{E}(r;t)$ as the set of extra classes that r connects to (or receives connections from) relative to those of t.

With the above definition, it can be readily seen that when $r \succ t$,

$$\bar{k}_r = \bar{k}_t + \sum_{v \in \mathcal{E}(r;t)} n_v, \tag{28}$$

where the index v runs over the extra classes to which r is connected. This implies that

$$\bar{k}_r^{\bar{k}_t} = \bar{k}_t^{\bar{k}_t} \left(1 + \frac{\sum_{v \in \mathcal{E}(r;t)} n_v}{\bar{k}_t} \right)^{\bar{k}_t}. \tag{29}$$

Thus we find that

$$q_{tr} = \begin{cases} \dfrac{\gamma_t \pi_t}{\bar{k}_t^{\bar{k}_t}} & r = t, \\ \dfrac{\gamma_t \pi_r}{\bar{k}_t^{\bar{k}_t}} \left(1 + \dfrac{\sum_{v \in \mathcal{E}(r;t)} n_v}{\bar{k}_t} \right)^{-\bar{k}_t} & r \succ t, \\ 0 & \text{o/w.} \end{cases} \tag{30}$$

[with $\mathcal{E}(t;t) \equiv \emptyset$]. Note that when $r \succ t$ and for large $\bar{k}_t$, q_{tr} deviates from our ansatz, Eq. (24), by an exponentially small amount.

Treating the deviations caused by the presence of cover relations among the classes as a small perturbation to our ansatz, Eq. (24), we obtain the leading order expression for q_{tr} as

$$q_{tr} = \begin{cases} 1 - \sum_{r \succ t} \dfrac{\pi_r}{\pi_t} \left(1 + \dfrac{\sum_{v \in \mathcal{E}(r;t)} n_v}{\bar{k}_t} \right)^{-\bar{k}_t} & r = t, \\ \dfrac{\pi_r}{\pi_t} \left(1 + \dfrac{\sum_{v \in \mathcal{E}(r;t)} n_v}{\bar{k}_t} \right)^{-\bar{k}_t} & r \succ t, \\ 0 & \text{o/w,} \end{cases} \tag{31}$$

where γ_t has been determined from the normalization

$$\sum_r q_{tr} = 1. \tag{32}$$

To the same order, we find also that

$$\pi_r = \frac{n_r}{N} - \sum_{t \succ r} \frac{n_t}{N} \left(1 + \frac{\sum_{v \in \mathcal{E}(t;r)} n_v}{\bar{k}_r} \right)^{-\bar{k}_r} + \sum_{t \prec r} \frac{n_r}{N} \left(1 + \frac{\sum_{v \in \mathcal{E}(r;t)} n_v}{\bar{k}_t} \right)^{-\bar{k}_t}. \tag{33}$$

Equations (31) and (33) are the analytical solution of the EM equations for a content-based network with connectivity function $c(r,s)$.

We see that whenever a class $r \succ t$, there is a nonzero probability for a node t to be also classified as belonging to class r. We will refer to this as a leakage in the class assignment. However, as can be seen from Eq. (31), the leakage probabilities vanish exponentially with the size of the classes with which t is connected. The more nodes (information) available in the system, the easier it is not to make mistakes in the classification of nodes of the covered class. A detailed account of the solution structure for content-based networks as well as more general types of networks will be given elsewhere [39].

When the content-based network is cover-free, the generalized EM equations have a leak-free solution and thus the entropy of the class assignments S_q vanishes. On the other hand, in the presence of cover relations, the EM method will produce assignments with some nodes in multiple classes, i.e., leaks. We have already found above the leading order behavior for the leakage. It is not too difficult to show that, in that case, S_q is given by

$$S_q = \sum_{t \text{ has a cover}} \sum_{r \succ t} n_r \alpha(r;t) \left(1 + \frac{\alpha(r;t)}{\bar{k}_t} \right)^{-\bar{k}_t}, \tag{34}$$

where $\alpha(r;t) \equiv \sum_{v \in \mathcal{E}(r;t)} n_v$ is the number of nodes to which nodes in class r are connected in addition to those that nodes in class t connect. In many practical situations, the number of contents is fixed. This implies that if the probability of being in class r is given by p_r, the actual number of nodes in the r class will grow on average as $n_r = N p_r$ with the system size. Therefore, the factors $\alpha(r;t)$ and $\bar{k}_t$ of Eq. (34) can also be written as

$$\alpha(r;t) = aN \quad \text{and} \quad \bar{k}_t = bN, \tag{35}$$

where a and b are constants whose values depend on the connectivity function that generated the network. Under these assumptions, the entropy S_q will decrease exponentially with the network size, meaning that even for moderately sized networks the leakages will be in general too small to cause significant misclassification.

As shown in Sec. IV, the solution of the EM equations provides us with an estimate for the class connectivity, $\tilde{c}(r,s)$, given by Eq. (21). For content-based networks without cover relation we have, *cf.* Eq. (22), $p(i|r) = \delta_{x_i,r}/n_r$, and

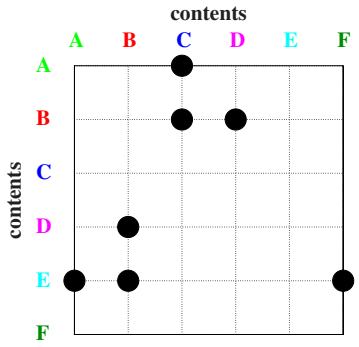

FIG. 3. (Color online) Connectivity function $c(x,y)$ for the theoretical example of Sec. V A. The number of contents is six: A,B,C,D,E, and F. The points represent the ones in the connectivity matrix, the values not marked are zero.

from Eq. (21) we find that $\tilde{c}(r,s)=c(r,s)$ with $S_c=0$. In the presence of cover relations among classes, there will be corrections vanishing exponentially with the number of nodes in the relevant classes. These results demonstrate that the EM algorithm is capable of inferring the hidden class connectivity function that generated the network.

A. Example

In order to further illustrate the theoretical results above, we turn next to an example. Consider a network generated from six kinds of contents to be denoted by A,B,C,D,E, and F, and with the connectivity function as shown in Fig. 3. The following cover relations are present: $B \succ A \succ F$; that is, $B \succ A$, $B \succ F$, and $A \succ F$. In fact, we have chosen this particular example to elucidate the effect of having nested covers and to show that the cover relation is transitive. For each of the cover relations, the sets of connections to additional classes are $\mathcal{E}(B;A)=\{D\}$, $\mathcal{E}(B;F)=\{D,C\}$ and $\mathcal{E}(A;F)=\{C\}$. When inserted into Eq. (31), these relations yield

$$q_{AA} = 1 - \frac{n_B}{n_A}\left(1 + \frac{n_D}{n_E + n_C}\right)^{-n_E - n_C},$$

$$q_{AB} = \frac{n_B}{n_A}\left(1 + \frac{n_D}{n_E + n_C}\right)^{-n_E - n_C},$$

$$q_{FF} = 1 - \frac{n_A}{n_F}\left(1 + \frac{n_C}{n_E}\right)^{-n_E} - \frac{n_B}{n_F}\left(1 + \frac{n_C + n_D}{n_E}\right)^{-n_E},$$

$$q_{FA} = \frac{n_A}{n_F}\left(1 + \frac{n_C}{n_E}\right)^{-n_E},$$

$$q_{FB} = \frac{n_B}{n_F}\left(1 + \frac{n_C + n_D}{n_E}\right)^{-n_E}, \tag{36}$$

with $q_{BB}=q_{CC}=q_{DD}=q_{EE}=1$ and all the other values of $q_{rt}=0$. These results are in agreement with what one would expect intuitively. For example, since $B \succ A$ and $B \succ F$, there is a nonzero probability of mistaking nodes of type A or F by nodes of B, i.e. q_{AB}, q_{FB}, and q_{FA} are all nonzero. However, these probabilities vanish exponentially with the number of nodes in classes E and C that are those with which the covered classes A and F have connections. In the large network size limit, the leakage on q_{ir}, and how far S_q deviates from zero, are determined by the pair of classes (r,t) such that r is the "tightest" cover of t, these are the pairs $r \succ t$ for which $\alpha(r;t)$ is smallest, *cf* Eq. (34): $A \succ F$, $B \succ A$ in our example.

VI. SIMULATION RESULTS: EM APPLIED TO CONTENT-BASED NETWORKS

In the following, we study numerically the ideas introduced in the previous sections. The generalized version of EM will be applied to directed content-based networks generated randomly from the connectivity functions shown in Fig. 4. The nodes of these networks have a content assigned that is selected at random out of $\mathcal{N}_x=5$, five possible contents, denoted by A,B,C,D, and E. Since the presence of coverage relations can change the quality of an EM classification, we have considered two connectivity functions $c(x,y)$ (see Fig. 4); one without class coverage, $c_{\mathcal{A}}$, and another, $c_{\mathcal{B}}$, with a single cover relation between contents A and B, such that $A \succ B$. In order to improve our numerical estimation of the classification with maximum likelihood, we implemented a simulated-annealing type of procedure for the optimization of $\bar{\mathcal{L}}$.

In the previous section, we have shown that our generalized EM method is able to infer the underlying content-based structure that generated the network. These calculations were carried out assuming that the number of contents $\mathcal{N}_x$ coincides with the number of classes $\mathcal{N}_c$. Let us therefore start by setting $\mathcal{N}_c=\mathcal{N}_x=5$. In Fig. 5, we show graphically the classifications obtained from the generalized EM method as applied to two networks of size $N=50$ generated with the connectivity functions of Fig. 4. The color coding is based on

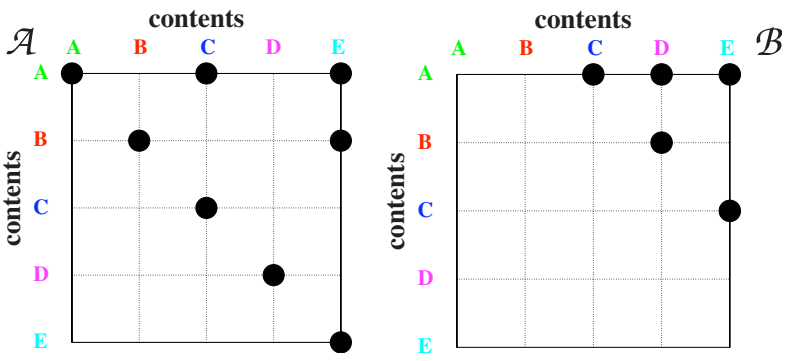

FIG. 4. (Color online) Connectivity functions $c(x,y)$ for the two examples of content-based networks analyzed in the simulation sections. The number of contents considered is five, A,B,C,D, and E. The contents of the connectivity function ($\mathcal{A}$) display no cover relation, while in the second example, ($\mathcal{B}$), $A \succ B$. The networks are generated assuming equal probability for the five contents at the assignation of a content to each node.

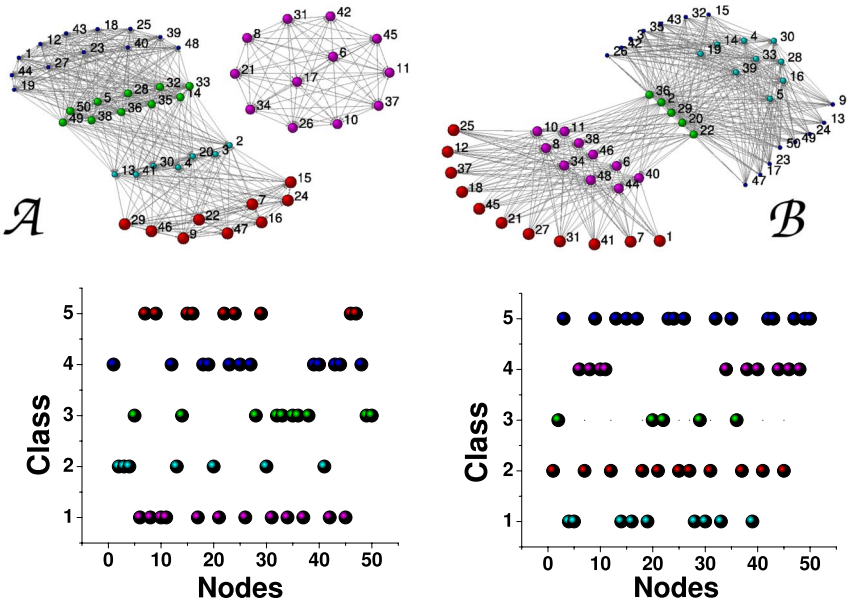

FIG. 5. (Color online) An example of classification, the original network is on the top and on the bottom the probability q_{ir} is represented graphically. The color and size of the symbols correspond to the contents of the nodes (green A, red B, magenta C, blue D, and cyan E). On the bottom, the radius of the spheres is proportional to the probability q_{ir}. The network $\mathcal{A}$ is generated using the connectivity function $c_{\mathcal{A}}$ of Fig. 4 with no cover relation among the classes, while on the right we have used $c_{\mathcal{B}}$, which incorporates a single cover relation between A and B such that $A>B$.

the contents of the nodes and will be such that it matches in all the subsequent figures of the paper (A green, B red, C blue, D magenta, and E cyan). The size of the spheres in the bottom plots are proportional to the probabilities q_{ir}. For these examples the classification is rather good even in the case when a cover relation is present, as can be readily seen from the bottom diagrams where no major color is misplaced. In other words, there are no misclassifications, although for the $\mathcal{B}$ case a slight amount of leakage (of order $\sim 10^{-6}$).

To try to quantify the quality of these results, we can, as a first measure, count the number of network realizations in our ensemble for which at least two nodes with different contents have been assigned to the same class, with the understanding that a node i is assigned to a class r whenever $q_{ir}>1/2$. This is a strict criterion, since it may well be that we are considering as *erroneous* a classification with only a single node misclassified. The result can also slightly depend on the method applied to optimize the likelihood. Still, this definition is a way to play on safe ground and avoid complicating too much the detection of mistakes in the classification. Let us call this then the error rate of the classification ϵ. For each of the two connectivity functions of Fig. 4, we have studied over 2000 realizations of networks of size $N=50$. In none of them the generalized EM method misclassified a single node. This result is in agreement with our earlier observation that the EM method classifies structurally equivalent nodes in the same way.

The next question is then: how can the optimal $\mathcal{N}_c$ be determined? If the networks studied are content based, there are several possible answers to this. Here we will outline two of them and will discuss at the end of this section a third one in the context of inferring the class connectivity function. In Sec. IV, we have introduced a measure S_q for the quality of an EM classification of the network. We have also shown that when $\mathcal{N}_x=\mathcal{N}_c$, S_q is either zero or exponentially small for large content-based networks. Therefore, a signal on S_q can be expected for $\mathcal{N}_c=\mathcal{N}_x$, if the EM algorithm is faced with the challenge of classifying a content-based network with a series of values $\mathcal{N}_c$. This effect happens because the normalization conditions of Eqs. (5) and (13) impose that no class can be left totally unassigned, $\pi_r>0$ for all r. The more redundant classes the method has to assign nodes to, the higher S_q becomes. In other words, we are providing the EM algorithm with a larger degree of freedom than required to properly classify the nodes. The extra freedom leads to structural leakage. The evolution of S_q with $\mathcal{N}_c$ is displayed in Fig. 6 for the two networks of Fig. 5. These are, of course, particular examples but some general features can be deduced. First, the value of S_q is rather small or even zero for $\mathcal{N}_c<\mathcal{N}_x$. This may be a generic property of content-based networks. As noted before, the structural equivalence of nodes with the same content prevents the EM algorithm from putting such nodes into different classes. This means that

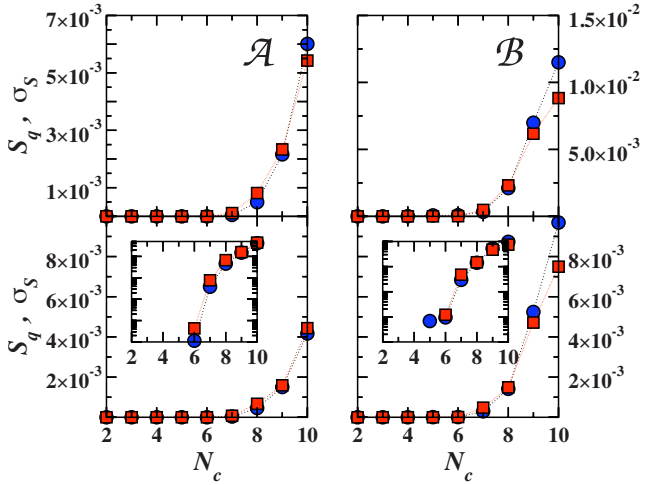

FIG. 6. (Color online) Shown in the lower panels are S_q (circles) and its fluctuations σ_S (squares) as a function of $\mathcal{N}_c$ for the networks of Fig. 5. In order to facilitate visualization, the insets show the same curves in a semilogarithmic plot. The top panels display the same quantities, S_q and σ_S, but ensemble averaged over different realizations of the content-based networks generated with the connectivity function of Fig. 4.

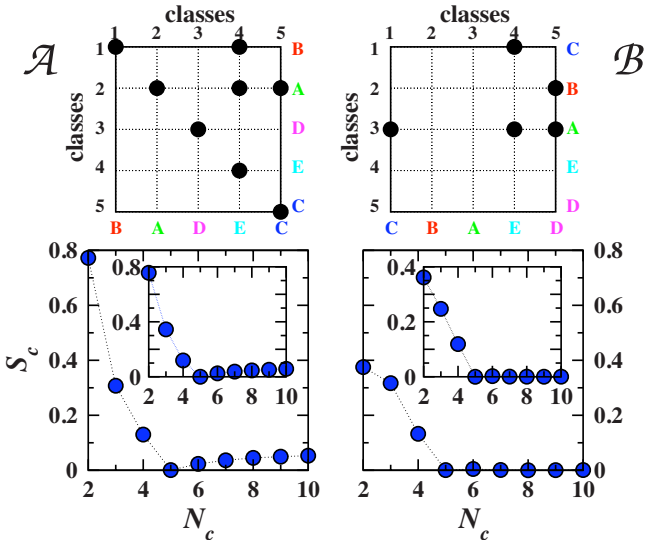

FIG. 7. (Color online) On the top, the connectivity function $\tilde{c}(r,s)$ obtained from the EM classification of the networks displayed in Fig. 5. The radii of the circles is proportional to the value of $\tilde{c}(r,s)$. On the bottom, we are showing how S_c goes with $\mathcal{N}_c$ for the same networks as well as, in the inset, an average over different content-based realizations generated with the connectivity functions of Fig. 4.

once the contents are classified by classes the leakage comes from cover relations between classes and can become very small for big networks. On the other hand, when $\mathcal{N}_c > \mathcal{N}_x$, the availability of excess classes that cannot be left totally unassigned causes S_q to be nonzero and to increase steadily with $\mathcal{N}_c$. The boundary between these two types of behaviors is precisely the unknown $\mathcal{N}_c = \mathcal{N}_x$.

Another peculiarity of the EM method applied to content-based networks is that when $\mathcal{N}_c < \mathcal{N}_x$, the landscape of the likelihood seems to have a very clear and unique maximum. The solution at the point of maximum $\bar{\mathcal{L}}(\pi,\theta)$ has also a well determined value of S_q. However, if $\mathcal{N}_c \geq \mathcal{N}_x$, the landscape of the likelihood becomes rough, with a large number of local maxima. The search for the global maximum under such conditions is therefore much harder. And even, in the cases where it can be numerically found, say when $\mathcal{N}_c = \mathcal{N}_x$, it is formed by a set of degenerate extrema with the same value of $\bar{\mathcal{L}}$ but very different values of S_q. Indeed, the values of the entropy shown in Fig. 6 for $\mathcal{N}_c \geq \mathcal{N}_x$ are averages over the best likelihood solutions found in different realizations of the optimization methods along with their standard deviations σ_S. The dispersion σ_S, of S_q around its average, can be used in practice as another estimator for the optimal number of classes (see Fig. 6).

Once $\mathcal{N}_x$ is known, it is possible to recover $c(r,s)$ as explained in Sec. IV. In the top panels of Fig. 7, the recovered $\tilde{c}(r,s)$ is displayed for the content-based networks of Fig. 5. After the classes of $\tilde{c}(r,s)$ have been properly reordered, it is impossible to distinguish the top panels of Fig. 7 from the connectivity functions given in Fig. 4. Also, in the lower panels of Fig. 7, we have included the evolution of the entropy S_c as a function of $\mathcal{N}_c$. S_c also shows a clear change of behavior at $\mathcal{N}_c = \mathcal{N}_x$, suggesting that the best content-based partition of the network happens when the number of classes equals the number of contents. Consequently, S_c, apart from being an estimator of how much a network deviates from a purely content-based graph, is also a useful quantity for deciding when $\mathcal{N}_c$ is optimal.

VII. EM AND NOISY CONNECTIONS IN CONTENT-BASED NETWORKS

It is unlikely that in real-world networks the generating processes is error-free. Even if the underlying structure is expected to be a content-based network, errors in the connecting pattern could naturally arise. We try to mimic the unexpected connections as well as the absence of expected connections, by introducing the corresponding error probabilities to the process of network generation from its contents. As before, each node i has a content x_i assigned at random from the set of possible contents (in the case of our example networks the same five possibilities: A, B, C, D, and E). Once the contents are established, the structure of the content-based network should be determined completely by the connectivity function $c(x_i,x_j)$: If $c(x_i,x_j)=1$, there ought to be a link from node i to j, and none if $c(x_i,x_j)=0$. As a way of gradually loosing the content-based structure of the connections, we introduce now the probabilities P_μ, and P_α, of not having a link, when $c(x_i,x_j)=1$ and having a link although $c(x_i,x_j)=0$, respectively. The networks constructed in this way can be regarded as hidden variable graphs [33–35] for which the probability of connection between any nodes i and j is expressed as

$$r(x_i,x_j) = c(x_i,x_j)(1-P_\mu) + [1-c(x_i,x_j)]P_\alpha. \qquad (37)$$

In other words, where in the absence of noise the probability of having a connection was one, it now is $1-P_\mu$, and likewise, where it was zero, it now is P_α. The extreme limit of this model occurs when $P_\mu = P_\alpha = 1/2$, so that the probability of connecting to a node of another class is maximally random and independent of the connectivity function. We are more interested here in the limit when both P_α and P_μ are much smaller than $1/2$, and the resulting graphs can be seen as a slight modification of a content-based network. For the sake of simplicity, all of the results shown below are for $P_\alpha = P_\mu \equiv P$.

Let us begin by looking at how the networks change with increasing assignment error. In the top panels of Fig. 8, we display a series of networks generated with the connectivity function $c_{\mathcal{A}}$ for different values of P. It is readily seen that the connection patterns associated with the different kinds of content becomes more and more diffuse. On the bottom panels of the same figure, we show the corresponding class assignment probabilities q_{ir}. While these are just examples, there are some features that are worth pointing out. The problems in the classification seem to appear somewhere between $P=1\%$ and $P=10\%$. Even at 10% of error the number of nodes misclassified in these networks is not very high. A closer inspection of the solution found shows that actually only two of the node content classes are mingled up, while all the remaining node classes are perfectly assigned. With the aim of quantifying these observations, the behavior of ϵ

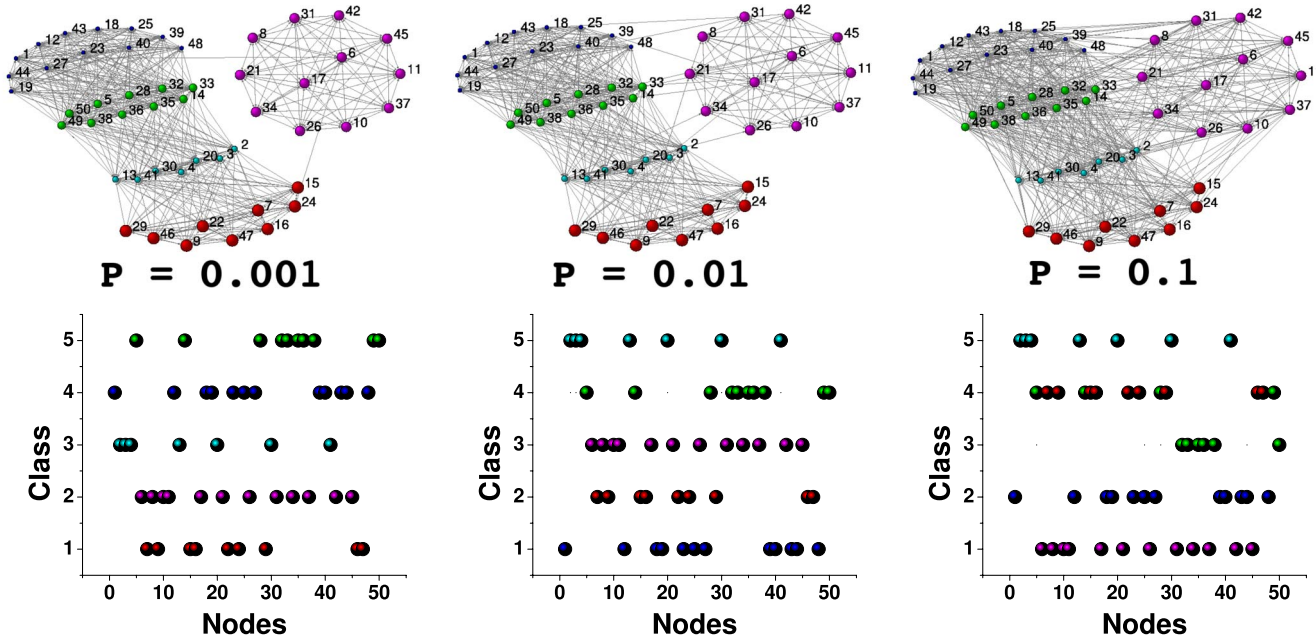

FIG. 8. (Color online) Same network as in section $\mathcal{A}$ of Fig. 5 but with increasing error probability P. The values of P are from left to right 0.001, 0.01, and 0.1. The plots on the lower panel are a graphic representation of the probability of classifying node i in class r, q_{ir}, as before the radius of the spheres are proportional to q_{ir} and the colors correspond to the actual content of the nodes (green A, red B, magenta C, blue D, and cyan E).

is plotted in Fig. 9 versus the disorder probability. This plot is, of course, susceptible to slight changes depending on the method used to search for the maximum likelihood and depends on how many realizations of the content-based graphs were considered (in this case 1000). Nevertheless, in our simulations the threshold for a sharp classification of all the nodes of the network is around $P \approx 7\%$ for graphs without coverage, connectivity function $c_{\mathcal{A}}$, and much lower, around $P \approx 5\%$, for those with a cover relation, $c_{\mathcal{B}}$. The exact value will depend on the particular connectivity function, apart from the optimization method, but these values give us already an idea about the order of magnitude of the threshold beyond which the content-based structure cannot be recovered anymore.

The next aspect to consider is how the entropies S_q and S_c are affected by the intensity of the disorder, and whether they are still valid estimators to determine the optimal number of classes. To answer this question, we fix the probability P at 1%, which seems to be a value where one might plausibly expect to obtain good classifications for both types of networks. In Fig. 10, we display S_q, σ_S, and S_c, as functions of the number of classes $\mathcal{N}_c$ with the results averaged over different content-based realizations. Indeed, at this level of disorder the entropies can still be used to estimate $\mathcal{N}_x$. The noise in the connections introduces a small constant background for S_c, which we will denote by S_c^*, and which can be determined in both examples from the behavior at high values of $\mathcal{N}_c$. We can estimate the value of S_c^* by noting that when $\mathcal{N}_c = \mathcal{N}_x$, any nonzero entropy should essentially be due to the background from the random connections.Substituting the expression for $r(x_i, x_j)$, Eq. (37), into the definition of S_c, Eq. (23), should therefore give us an estimate for S_c^*,

$$S_c^* \approx -\frac{2}{\mathcal{N}_c^2 \ln 2} \sum_{x,y} r(x,y) \ln r(x,y). \tag{38}$$

For $P=1\%$, this yields $S_c^* \sim 0.112$, close to the value observed in the Fig. 10 for $\mathcal{N}_c \geq 5$. To check how well our estimate for S_c^* agrees with the values obtained from simulations, we plot in Fig. 11 S_c vs the disorder probability at $\mathcal{N}_c=5$. When the disorder becomes very strong, on the other hand, it might not be possible to find an optimal $\mathcal{N}_c$. Moreover, the presence of very different connection patterns for nodes with the same content renders the existence of such an optimal number dubious. Therefore, apart from the obvious classification $\mathcal{N}_c=N$, there may not be any other sharp classification. The effects of high disorder can be seen in Fig. 10, where the entropies S_c and S_q are represented as functions of

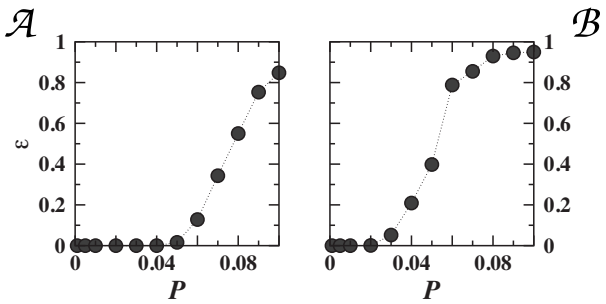

FIG. 9. The error rate ϵ as a function of the error probability P for content-based networks generated with the connectivity functions of Fig. 4 and with $\mathcal{N}_c=\mathcal{N}_x=5$.

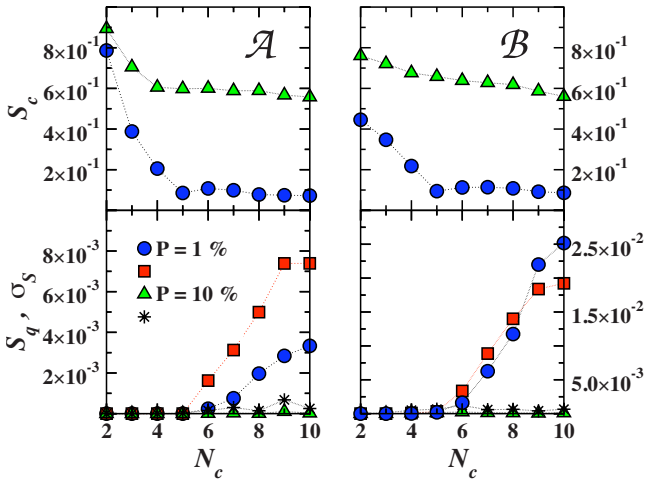

FIG. 10. (Color online) The average entropies over different realizations for content-based networks generated with the connectivity functions of Fig. 4. In the top panels, S_c is represented as a function of the number of classes $\mathcal{N}_c$ for two different levels of disorder: the circles are $P=1\%$, while the triangles for $P=10\%$. On the bottom panels, S_q and σ_S versus $\mathcal{N}_c$ for the disorder probabilities $P=1\%$, circles (S_q), and squares (σ_S), and $P=10\%$, triangles (S_q), and stars (σ_S).

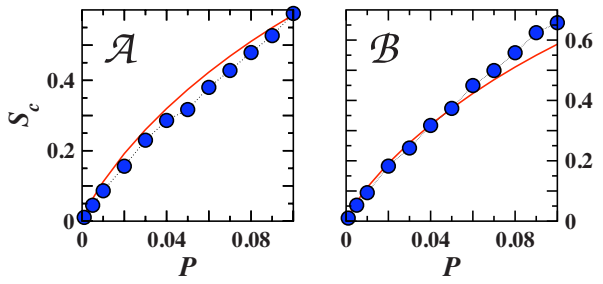

FIG. 11. (Color online) The average entropy S_c as a function of the disorder probability P for content-based networks generated with the connectivity functions $c_{\mathcal{A}}$ and $c_{\mathcal{B}}$ depicted in Fig. 4. The red curves correspond to the value of S_c^*.

$\mathcal{N}_c$ for $P=10\%$. The results depend on the connectivity function, $c_{\mathcal{A}}$ seems a little more robust to the disorder as was confirmed by Fig. 9, but the signal in S_q or σ_S is clearly lost or has moved to higher values of $\mathcal{N}_c$. Also, S_c has lost its capacity to predict $\mathcal{N}_x$ and smoothly falls for higher and higher values of $\mathcal{N}_c$.It is worth noting that in spite of the lack of a method to find $\mathcal{N}_x$, if $\mathcal{N}_c=5$, the EM method retrieves the appropriate hidden variable theory connectivity function $r(x,y)$ as can be inferred from the good fit produced by Eq. (38) to S_c shown in Fig. 11.

The numerical findings of this section show that the classifications of the EM method are robust to the introduction of noise in the connection patterns up to a certain point. The certainty of the classification will suffer, the stronger the disorder becomes. In fact this is one of the major merits of the EM method: it is able to extract the underlying content-based structure even in the presence of a certain level of noisy connections.

VIII. CONCLUSION

In summary, we have shown how the EM method for the classification of nodes of a network can be applied to content-based networks in order to extract the underlying content-based structure even in the presence of a certain level of disorder in the connections. The application of the EM method to content-based networks is a natural concept that follows from the observation that the EM method classifies structurally equivalent nodes in an identical manner. In this sense, the EM method can be related to the block modeling techniques proposed in the social sciences. Content-based networks, on the other hand, are of great relevance, since they can be regarded as idealized paradigms of networks with communities or multipartite structures, including mixtures of both. Since in many realistic graphs the vertices carry additional attributes which might influence or even determine their connections to other vertices, being able to extract any content-based pattern can provide information about how the networks emerged.

Our approach in this paper has been to start out with pure content-based graphs, and to show analytically as well as numerically that the EM method can infer the content-based connectivity pattern. We have shown also that the existence of cover relations between contents leads to nonzero probabilities of mistaking nodes belonging to different classes. However, these probabilities vanish exponentially with the increasing number of nodes, i.e., the more discriminating information provided to the method. By regarding more realistic networks as perturbations of content-based networks under the addition or removal of connections, we then asked under which circumstances the EM method is still able to perform satisfactorily. There is a certain level of disorder beyond which the inference of the content-based structure, specially the number of contents, becomes rather difficult if not impossible.

In order to estimate the quality of the classification and how far the structure of the network is from a content-based structure, we have introduced two entropies, S_q and S_c, which actually can be useful for the classification of any kind of graphs, including real-world networks. We have also shown that these entropies are applicable to deduce the optimal number of classes needed by the EM method to obtain a sharp classification of the nodes of the network.

ACKNOWLEDGMENTS

The authors would like to thank Alessandro Vespignani, Santo Fortunato, Filippo Radicchi, and in general the members of the Cx-Nets collaboration for useful discussions and comments. Funding from the Progetto Lagrange of the CRT Foundation, the Research Fund of Boğaziçi University, as well as the Nahide and Mustafa Saydan Foundation was received. In addition, M.M. would like to acknowledge the kind hospitality of the ISI Foundation.

[1] P. Doreian, V. Batagelj, and A. Ferligoj, *Generalized Blockmodeling* (Cambridge University Press, Cambridge, 2005).
[2] L. C. Freeman, *The Development of Social Network Analysis* (Empirical, Vancouver, 2004).
[3] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási, Science **297**, 1551 (2002).
[4] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[5] M. Grivan and M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).
[6] F. Radicchi *et al.*, Proc. Natl. Acad. Sci. U.S.A. **101**, 2658 (2004).
[7] S. Fortunato and M. Barthélemy, Proc. Natl. Acad. Sci. U.S.A. **104**, 36 (2007).
[8] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004); Proc. Natl. Acad. Sci. U.S.A. **103**, 8577 (2006).
[9] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **103**, 8577 (2006).
[10] S. Fortunato and C. Castellano, in *Encyclopedia of Complexity and System Science* (Springer, Berlin, 2008).
[11] T. I. Lee *et al.*, Science **298**, 799 (2002).
[12] B. Alberts *et al.*, *Molecular Biology of the Cell* (Garland Science, New York, 2002), Chap. 9.
[13] M. M. Babu *et al.*, Curr. Opin. Struct. Biol. **14**, 283 (2004).
[14] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[15] A. Arenas, A. Díaz-Guilera, and C. J. Pérez-Vicente, Phys. Rev. Lett. **96**, 114102 (2006).
[16] S. Boccaletti, M. Ivanchenko, V. Latora, A. Pluchino, and A. Rapisarda, Phys. Rev. E **75**, 045102(R) (2007).
[17] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, Phys. Rev. E **70**, 025101(R) (2004).
[18] J. Reichardt and S. Bornholdt, Phys. Rev. Lett. **93**, 218701 (2004); Phys. Rev. E **74**, 016110 (2006); Physica D **224**, 20 (2006).
[19] J. S. Kumpula, J. Sarämaki, K. Kaski, and J. Kertész, Eur. Phys. J. B **56**, 41 (2007).
[20] G. Palla *et al.*, Nature (London) **435**, 814 (2005); G. Palla, A.-L. Barabási, and T. Vicsek, *ibid.* **446**, 664 (2007).
[21] H. Zhou, Phys. Rev. E **67**, 041908 (2003); **67**, 061901 (2003).
[22] I. Simonsen *et al.*, Physica A **336**, 163 (2004).
[23] D. Gfeller, J.-C. Chappelier, and P. De Los Rios, Phys. Rev. E **72** 056135 (2005).
[24] F. Lorrain and H. C. White, J. Math. Sociol. **1**, 49 (1971).
[25] H. C. White, S. A. Boorman, and R. L. Breiger, Am. J. Sociol. **81**, 730 (1976).
[26] D. Balcan and A. Erzan, Eur. Phys. J. B **38**, 253 (2004).
[27] M. Mungan, A. Kabakçioğlu, D. Balcan, and A. Erzan, J. Phys. A **38**, 9599 (2005).
[28] D. Balcan, A. Kabakçioğlu, M. Mungan, and A. Erzan, PLoS ONE **2**, e501 (2007).
[29] D. Balcan and A. Erzan, Chaos **17**, 026108 (2007).
[30] M. E. J. Newman and E. A. Leicht, Proc. Natl. Acad. Sci. U.S.A. **104**, 9564 (2007).
[31] J. D. Farmer, N. H. Packard, and A. S. Perelson, Physica D **22**, 187 (1986).
[32] A. S. Perelson and G. Weissbuch, Rev. Mod. Phys. **69**, 1219 (1997).
[33] G. Caldarelli, A. Capocci, P. De Los Rios, and M. A. Munoz, Phys. Rev. Lett. **89**, 258702 (2002).
[34] B. Söderberg, Phys. Rev. E **66**, 066121 (2002).
[35] M. Boguñá and R. Pastor-Satorras, Phys. Rev. E **68**, 036112 (2003).
[36] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions* (Wiley, New York, 1996).
[37] D. Garlaschelli and M. I. Loffredo, e-print arXiv:cond-mat/0609015.
[38] The equations written in this form take care of the case when some of the θ are zero, as can be readily checked by comparing their solution, Eq. (18), with Eqs. (12) and (16).
[39] M. Mungan and J. J. Ramasco (unpublished).