Correlation clustering on networks

2009 J. Phys. A: Math. Theor. 42 345003

(http://iopscience.iop.org/1751-8121/42/34/345003)

IOP PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND THEORETICAL

J. Phys. A: Math. Theor. **42** (2009) 345003 (15pp) doi:10.1088/1751-8113/42/34/345003

Correlation clustering on networks

Z Néda[1], R Sumi[1,2], M Ercsey-Ravasz[3], M Varga[1], B Molnár[1] and Gy Cseh[1]

[1] Department of Theoretical Physics, Babeş-Bolyai University, Str. M. Kogălniceanu 1, RO-400084, Cluj-Napoca, Romania
[2] E-Austria Research Institute, Bd. V. Pârvan 4, RO-300223, Timişoara, Romania
[3] Department of Physics, University of Notre Dame, Center for Complex Network Research, Notre Dame, IN-46556, USA

E-mail: zneda@phys.ubbcluj.ro

Received 8 April 2009, in final form 15 June 2009
Published 6 August 2009
Online at stacks.iop.org/JPhysA/42/345003

Abstract

Random networks with co-existing positive and negative links are studied from the viewpoint of the NP hard correlation clustering problem. The task is to produce a clustering of the vertices which maximizes the number of positive edges within clusters and the number of negative edges between clusters. Simulated annealing, Monte Carlo renormalization and molecular dynamics optimization are used to find the optimal cluster structure. Recently, this problem was studied for globally coupled systems and an interesting phase-transition-like phenomenon was predicted: in the thermodynamic limit the relative size of the largest cluster, r, exhibits a step-like behavior as a function of the density of positive links q ($r = 0$ if $q < 1/2$ and $r = 1$ if $q > 1/2$). Here we prove that when considering random networks with a constant bond density, the same phase transition is expected. A totally different result emerges however, when networks with a fixed average number of connections per node are considered. In such cases a nontrivial spin-glass-type behavior is found, where the location of the critical point shifts toward $q > 1/2$ values. The results also suggest that instead of the simple step-like behavior, the $r(q)$ curve has a more complex shape, which depends on the specific topology of the considered network.

PACS numbers: 64.60.aq, 75.10.Nr, 05.10.Ln, 83.10.Rs

1. Introduction

Correlation clustering (CC) [1–3] is an NP hard optimization problem with potential applications in computer science, physics, sociology, psychology, information technology and medicine. This problem can be formulated in layman's terms: given a set of vertices

globally connected through randomly chosen positive or negative links the goal is to find a clustering of them which maximizes the number of positive links within the clusters and the number of negative links between the clusters. Based on our daily-life experience, an intuitive sociological formulation is also possible: given a set of agents with symmetrical positive and negative propensities toward each other, find an optimal grouping of them so that the fixed propensities are optimally satisfied. This means that agents connected with positive propensities (agents that 'like' each other) should be in the same group while agents connected by negative propensities (agents which 'hate' each other) should be in different groups. One can easily realize that a perfect solution is usually not possible, since there is no optimal grouping which would satisfy all connections. A simple example in this sense is a 'frustrated triangle': three agents interconnected by two positive links and one negative link.

For a globally coupled system the problem is similar to the well-known Sherrington–Kirckpatrick (SK) spin-glass problem [4]. In the SK model, global interaction is assumed between Ising-like spins. The values of these interactions are randomly chosen, both positive and negative values being allowed, leading to an obvious frustration in the system. At $T = 0$ thermodynamic temperature a complex and computationally difficult task is to determine the spin-configuration with minimal energy. Computationally CC is even harder than the SK energy minimization, since in the SK problem one needs to divide the spins into two groups (spins up or down), while in the CC problem the number of clusters is also a variable that has to be optimized.

It is easy to see that the CC problem is relevant to many practical situations. It was originally motivated by research at Whizbang labs, where learning algorithms were trained to help various clustering tasks [5]. CC is also related to agnostic learning [6], which is an emerging approach to efficient data mining and artificial intelligence. An important application can be in medicine and pharmaceutics, where one needs to divide drugs into compatibility groups. Closely related problems were considered also while studying coalition formation phenomena in sociological systems [7–9]. The reason why the problem is interesting to the physics community is that it resembles the infinite-range p-state Potts-glass [10–14] and exhibits a phase transition-like phenomenon [15].

The CC problem can be formulated mathematically rigorously by introducing a K cost-function [15] which increases by $1/N$ (N is the number of agents in the system) whenever two conflicting agents are in the same cluster or when two agents with positive propensities between them are in different clusters:

$$K = \sum_{i<j} \frac{|J_{ij}| + J_{ij}}{2}(1 - \delta_{\sigma(i)\sigma(j)}) + \sum_{i<j} \frac{|J_{ij}| - J_{ij}}{2}\delta_{\sigma(i)\sigma(j)}. \tag{1}$$

In (1) $\sigma(i)$ denotes the cluster to which agent i belongs, the sums are for all possible pairs, δ_{ij} is the Kronecker delta symbol and $J_{ij} = \pm 1/N$ is the link between agent i and j. Simple algebra leads to a much simpler form of the K cost function:

$$K = -\sum_{i<j} \delta_{\sigma(i)\sigma(j)} J_{ij} + \frac{1}{2}\sum_{i<j}(J_{ij} + |J_{ij}|). \tag{2}$$

Solving the CC problem is equivalent to minimizing the K cost function, which thus represents a kind of energy (or Hamiltonian) for the system. In order to keep the analogy with thermodynamic systems the cost function has to be extensive, and its average value should scale linearly with system size. This is the reason why the J_{ij} interaction parameters are taken proportional to $1/N$. One immediately realizes that for a given distribution of the links the

second sum in (1) is constant and one has to minimize the simpler cost function

$$K = -\sum_{i<j} \delta_{\sigma(i)\sigma(j)} J_{ij}, \tag{3}$$

which resembles the Hamiltonian of the well-known infinite-range and infinite-state Potts glass [10–14]. The Potts-glass problem is different however in many senses from the CC problem. As discussed in a recent work [15] both the degeneracy of the states and the nature of the disorder are different for the two problems. Crucial for us is the statistics of the J_{ij} interactions which determines the nature of the disorder. For the Potts glass problem the disorder is relevant. This means that the J_{ij} interactions are defined as

$$\langle J_{ij}\rangle_{\text{Potts}} = \frac{J_0}{N}, \qquad (\Delta J_{ij})_{\text{Potts}} = \left\langle J_{ij}^2\right\rangle - \langle J_{ij}\rangle^2 \propto \frac{1}{N}. \tag{4}$$

The variance scales as a function of the system size in a similar manner with the mean, leading to a relevant disorder in the thermodynamic limit. This makes the system a complex one and leads to the observed spin-glass-type behavior.

For the classical CC problem however, the situation is simpler in the thermodynamic limit. To see this, let us denote the density (probability) of positive links by q ($q \in [0, 1]$). For a fully connected system this means that $J_{ij} = +1/N$ with probability q and $J_{ij} = -1/N$ with probability $1 - q$, leading to

$$\langle J_{ij}\rangle_{\text{CC}} = \frac{2q-1}{N}, \qquad (\Delta J_{ij})_{\text{CC}} = \frac{4q(1-q)}{N^2} \propto \frac{1}{N^2}. \tag{5}$$

The disorder is much weaker in this case, since the variance scales as $1/N^2$ in comparison with the mean that scales again as $1/N$. This means that the disorder scales out in the thermodynamic limit, and as a consequence the system behaves in a much simpler manner. As discussed in [15], in the thermodynamic limit an acceptable approximation is to replace all J_{ij} interactions with their mean value $\langle J_{ij}\rangle_{\text{CC}} = (2q-1)/N$. The solution of this problem is quite simple however. Whenever $\langle J_{ij}\rangle$ is positive ($q > 1/2$) the optimal solution is to put all agents in the same cluster and whenever $\langle J_{ij}\rangle$ is negative ($q < 1/2$) put all agents in separate clusters.

Although the perspectives for a simple solution were quite gloomy in the beginning one can see that the solution becomes simple in the thermodynamic limit. Before getting too excited about this, let us remember that all practically interesting cases are for finite N values, where the problem remains NP hard [16]. In such cases the best we can do is to consider some numerical optimization techniques such as simulated annealing, analytical or numerical renormalization approach or some other numerical optimization tricks [15, 17].

From the viewpoint of statistical physics, the CC problem becomes interesting due to the phase-transition-like behavior of the optimal cluster structure as a function of the q density of positive links. For a globally connected system this critical point is at $q = 0.5$. A proper order parameter, suitable for characterizing this transition is the relative size of the largest cluster. Rigorously, this order parameter is defined as

$$r(q) = \left\langle \left\langle \max_{(i)} \left\{ \frac{C_x\{i,q\}}{N} \right\} \right\rangle_{\text{deg}} \right\rangle_x, \tag{6}$$

where $C_x\{i, q\}$ denotes the number of agents in cluster i, for an x realization of the disorder (distribution of the J_{ij} interactions) with a fixed q density of the positive links. Since the ground state could be degenerated (a different cluster structure with the same minimum K value might exist), first an average over all these degenerated states is considered. Then, a second average over the quenched disorder x is performed. For finite system sizes, N, one can

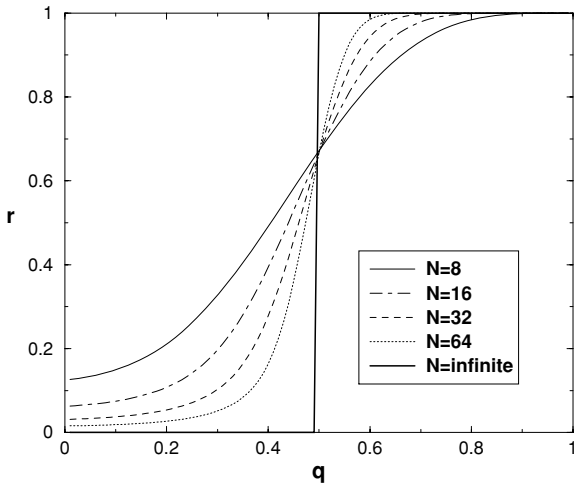

Figure 1. Analytical renormalization results for the $r(q)$ curve in the case of a globally coupled system.

compute the $r(q)$ curves using different numerical or analytical approximation techniques. A simple analytical renormalization technique, simulated annealing, extreme optimization and a molecular dynamics approach were considered [15, 17]. All these methods yield a picture which is consistent with our prediction for the $N \to \infty$ thermodynamic limit. As an example in figure 1 we present the analytical results of a simple renormalization approach (for details see [15]).

Globally coupled networks are seldom relevant to large natural and sociological systems [18–20]. An immediate question which arises then is to consider the CC problem on different random graph (network) structures. In such cases the majority of links between vertices will be absent, and the existing links will be positive with probability q and negative with probability $1 - q$. Particularly, we are interested in the shape of the $r(q)$ curves as a function of dilution (proportion of absent links), network size and topology.

2. Optimization methods

For finite system sizes finding the optimal cluster structure in the CC problem is a complex NP hard optimization problem. As it will be proved later, for strongly diluted random networks even the thermodynamic limit is NP complex (the disorder remains relevant). Numerical optimization techniques are used to study such systems and to compute the shape of the $r(q)$ curves.

Due to the fact that the disorder is quenched both in the distribution of the positive and negative links and in the realization of a particular link topology, the order parameter, r, is now an average over three different ensembles:

$$r(q) = \left\langle \left\langle \left\langle \max_{(i)} \left\{ \frac{C_x\{i,q\}}{N} \right\} \right\rangle_{\text{deg}} \right\rangle_x \right\rangle_{\text{net}} . \tag{7}$$

In contrast with (6), the third average here is on the particular realization of the network (denoted by the index 'net'). This average is calculated by generating different graphs with similar topological properties.

In the present work three different optimization techniques are used to determine the shape of the $r(q)$ curves. In the following, these methods are described and discussed in a critical manner.

2.1. Simulated annealing

Simulated annealing is implemented in the standard fashion [21].

(i) First the random network with N vertices is build by specifying the connectivity matrix (links).
(ii) The existing links are assigned positive or negative values (+1 or −1) randomly, respecting the q probability of the positive links.
(iii) Initially all vertices (agents) are in different clusters and an initial T_i temperature is considered. (The temperature scale is defined in such way that the value of the Boltzmann constant is chosen as unity $k = 1$.) The value of this temperature will be consecutively lowered during the annealing algorithm.
(iv) For a given temperature many Monte Carlo (MC) steps are considered. One MC step is defined as N elementary simulation steps. In one elementary simulation step we randomly choose an agent, and reassign it to a randomly chosen cluster. This change is accepted with probability 1 if the cost function (energy) is lowered by that change ($\Delta K < 0$) and with probability $\exp(-\Delta K/T)$ if the cost function is increased by the change.
(v) After a desired number of MC steps are made for a temperature (usually this is of the order of 1000), the temperature is lowered with a constant rate: $T_{\text{new}} = 0.98T_{\text{old}}$, until a fixed final temperature, T_f, is reached. For the optimizations performed in the present work we considered $T_i = N/[-4\log(0.8)]$ and $T_f = 0.1$. Once T_f is reached the optimization is done and the relative size of the largest cluster is recorded.
(vi) For a given link distribution the whole optimization process is repeated several times (in our case 10 times).
(vii) Keeping the value of q and the particular network topology we reassign the $+/-$ links and perform another set of optimization. This second average, which is realized over the positive and negative links distribution is done again 10 times.
(viii) Finally, an average over the particular realization of the random net is done. This averaging is done by generating again 10 different networks.

The final average is thus a result of a modest average over three different ensembles and it is presumed to be independent of the particular realization of the quenched disorder. Since this averaging is computationally time consuming, only relatively small graphs, up to $N = 100$ vertex points, could be studied by this method.

2.2. Stochastic renormalization

The simple analytical renormalization method used for globally coupled systems in [15] motivated this heuristic method. The advantage of this method is that it is easy to implement and computationally it is less demanding than simulated annealing. In consequence, much larger graphs can be studied and surprisingly the results are in excellent agreement with the better established results of simulated annealing.

The basic idea is that for a network with only negative links the optimal cluster structure is known: each node should be in a different cluster. Starting from this simple configuration the links in the graph are visited in random order and turned to positive values. After each change we try to keep the cost function minimal by changing the cluster to which the involved

nodes belong. The q density of positive links is monotonically increased as more and more links are turned positive. For each q value the relative size of the largest cluster is recorded. Repeating the procedure several times (of the order of thousands), and considering a second average over the graph structure the same average order parameter as the one obtained in simulated annealing (7) is computed. The detailed steps of the algorithm are the following:

(i) Generate a graph with the needed statistical properties.
(ii) Initially all links are considered negative and all N nodes are placed in different clusters.
(iii) At each simulation step we randomly choose one of the negative links (J_{ij}) and turn it positive. If vertices i and j are already in the same cluster, repeat this step by choosing another link.
(iv) We compute the cost function of the system in such a case, K_{init}.
(v) The two nodes (i and j) belonging to this link are tentatively joined in a new cluster. All nodes belonging to clusters $\sigma(i)$ and $\sigma(j)$ are then joined in this cluster. The cost function, K_{fin}, is computed for the new cluster structure.
(vi) If $K_{\text{fin}} < K_{\text{in}}$ we accept the new cluster, otherwise the old cluster structure is restored.
(vii) If the new cluster structure is accepted we check all nodes belonging to this new cluster. If a node is found so that the cost function would further decrease if this node would be in its old cluster we put it back in its old cluster. This check is repeated until no such node is found.
(viii) This ends one simulation step.
(ix) The relative size of the largest cluster is recorded after each completed simulation step (this value belonging to a given q value of positive links).
(x) Steps 3–8 are repeated until all links are turned positive.
(xi) Steps 2–9 are repeated several thousands times to get a reasonable average.
(xii) We generate several new graphs with the needed statistical properties and steps (i)–(xi) are repeated for each of them.

The method is much faster than simulated annealing, and a reasonable average can be done on system sizes up to $N = 500$ nodes. For small network sizes one can compare the results of this method with those given by simulated annealing. The good agreement between these results offered a confidence for the applicability of this otherwise heuristic method.

2.3. Molecular dynamics approach

The molecular dynamics (MD) approach to CC is straightforward [22]. Let us consider N particles (the nodes) placed on a ring with unit radius ($R = 1$), and interacting through forces that decay exponentially as a function of the inter-particle separation angle. For characterizing the position of the particles we use polar coordinates. Since $R = 1$, the position of each particle is characterized solely by an angle Ψ_i expressed in degrees ($\Psi_i \in [0, 360]$). The force f_{ij} acting between two agents i and j is considered to be proportional to the strength of the J_{ij} link

$$f_{ij} = C J_{ij} \exp(-k x_{ij}), \tag{8}$$

where k and C are fixed positive constants, $J_{ij} = \pm 1$ and

$$x_{ij} = \min\{|\Psi_i - \Psi_j|, 360 - |\Psi_i - \Psi_j|\}. \tag{9}$$

For $f_{ij} > 0$ the interaction forces are attractive and for $f_{ij} < 0$ the forces are repulsive ones. Since each of these forces can act either in the positive or negative direction, the force

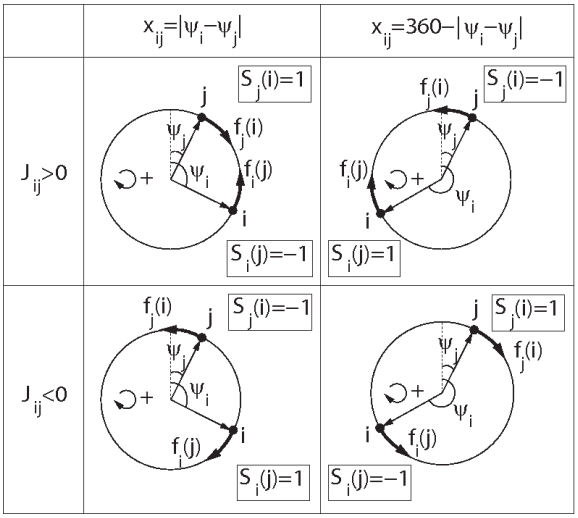

Figure 2. Forces between the elements for the possible configurations. The values of the $s_i(j)$ coefficients are given for each case in part.

has to be oriented. This can be done simply by taking into account the relative position of the two particles on the circle. The force $f_i(j)$ acting on particle i on behalf of particle j is

$$f_i(j) = s_i(j) f_{ij}, \tag{10}$$

where $s_i(j) = -1$ if $x_{ij} = \Psi_i - \Psi_j$ or $x_{ij} = 360 - (\Psi_j - \Psi_i)$, otherwise $s_i(j) = +1$. Figure 2 illustrates the possible cases and the values of the $s_i(j)$ coefficients. In each time moment, the resultant force, F_i, acting on each agent is computed:

$$F_i = \sum_{\{j\}, j \neq i} f_i(j). \tag{11}$$

Agents will follow an over-damped motion with parallel update [17, 22] of their position. At each dt time step a $d_i = F_i\,\mathrm{d}t$ displacement is considered for each agent. Particles are considered to be point-like and transparent, they can freely pierce through each other and an arbitrary number of them can occupy the same position in space. Due to the presence of the attractive and repulsive forces the system is in general strongly frustrated and the energy landscape is complex with several deep local minima. As a result of the simple molecular dynamics approach the system relaxes to one of its local equilibria, which of course might not be the optimal (global) one. The system is considered to be relaxed when the maximal displacement of the agents is less than a fixed value (considered here as 0.01°). In order to make the local minimum unstable, after relaxation random $F_{\mathrm{rand}}(i)$ forces are applied on each particle. The strength of these forces is uniformly distributed on an $F_{\mathrm{rand}}(i) \in \left[-F_{\mathrm{rand}}^{\max}, F_{\mathrm{rand}}^{\max}\right]$ interval. The purpose of this random force is to shake up the system and to make the local equilibrium configurations unstable. Repeating the relaxation procedure many times, the desire is to freeze the system in the global energy minimum. After one hundred of such relaxation steps we assume that the global minimum is reached and at this point the order parameter is computed. The optimization procedure applied here is similar in some sense with simulated annealing, random forces replacing the effect of the heat bath. The main difference relative to simulated annealing is that in the present algorithm the noise is not constantly applied and the intensity of the noise is kept constant.

The result of the molecular dynamics optimization is a spatial configuration of the agents on the unit circle. In order to determine the needed order parameter, r, one still has to identify the clusters. This is done in the following manner. We choose a small $\psi(1) = 1$ angle. Rows of consecutive neighboring agents separated by an angle smaller than $\psi(1)$ are assigned to the same cluster. We check whether the elements of the clusters are connected through the original network structure. If not, we break these clusters in their largest connected components. We calculate the cost function, $K(1)$, for the obtained clustering, and increase $\psi(i)$ with a $\mathrm{d}\psi = 1$ step. For $\psi(i+1) = \psi(i) + \mathrm{d}\psi$ we repeat the same clustering procedure and calculate $K(i+1)$. The value of ψ is consecutively increased until $\psi = 180$ is reached. The minimal $K(i)$ value will determine the optimal cluster structure.

3. Considered networks

Two categories of networks are considered: (i) completely random Erdős–Rényi type graphs [23, 24] and (ii) scale-free networks [18, 19]. Since all the previously discussed stochastic optimization methods are computationally time consuming and a complicated average is necessary, only relatively modest networks (up to $N = 500$ nodes) could be studied.

As a first step, for each generated graph one has to check whether it is connected, i.e. it does not fall in non-connected components. In case the graph splits in non-connected components a new graph has to be generated, since the solution of the CC problem is obviously different on a non-connected graph. Before presenting the results of the optimization methods on finite graphs let us discuss what would one expect in the thermodynamic limit, and how these graphs can be generated for a computational study.

3.1. Randomly diluted networks

Erdős–Rényi networks are simple random graphs on which many exact results are known [25]. A method to generate such graphs is by randomly diluting a fully connected graph: all links are visited and taken out with a fixed $(1 - p)$ probability. By this procedure one obtains a random graph with $pN(N-1)/2$ links on average and with a normal degree distribution. For studying the CC problem the remaining links will be considered $+1$ with a q probability and -1 with $(1 - q)$ probability. For a finite dilution probability one would expect again a simple solution for the CC problem. Similarly with the case of globally coupled graphs the disorder is irrelevant in the thermodynamic limit. To prove this, let us view the diluted graphs as fully connected ones with S_{ij} links, $S_{ij} = 0$ with $1 - p$ probability and $S_{ij} = \pm J \neq 0$ with p probability ($S_{ij} = J > 0$ with probability q and $S_{ij} = -J < 0$ with probability $1 - q$). In order to have the K cost-function extensive, its average value, $\langle K \rangle$, should scale linearly with the size of the system, N:

$$\langle K \rangle = JN(N-1)p(2q-1)/2 \propto N. \tag{12}$$

Now, to satisfy this scaling one needs to choose $J \propto (1/N)$. Taking the proportionality factor 1, it results

$$\langle S_{ij} \rangle_{\mathrm{CC}}^{\mathrm{ER}} = \frac{p(2q-1)}{N} \propto \frac{1}{N}, \qquad (\Delta S_{ij})_{\mathrm{CC}}^{\mathrm{ER}} = \frac{p - p^2(2q-1)^2}{N^2} \propto \frac{1}{N^2}. \tag{13}$$

Since the variance converges quicker to zero than the average, the disorder behaves as in the case of the globally coupled systems, i.e. it scales out in the thermodynamic limit. In this limit the system is equivalent to a system having all links $\langle S_{ij} \rangle = Jp(2q-1)$ and consecutively the same phase transition is expected at $q_c = 0.5$ as in the case of the globally

coupled systems. The critical point is thus independent of the considered dilution, assuming of course that p is finite.

3.2. Random networks with fixed average degree

Randomly diluted complete graphs with fixed dilution densities are still too dense for giving a non-trivial solution of the CC problem in the thermodynamic limit. Based on this observation, a further possibility is to consider Erdős–Rényi-type random graphs where the average degree (average number of links per node), $\langle k\rangle$, is fixed. This would result in infinite dilution for $N \to \infty$ since the dilution is increased as the system size increases. The average value of the cost function scales as $K = JN\langle k\rangle/2$, and will behave like an extensive thermodynamic function if we choose $J = C$. Taking $C = 1$, the network can be viewed again as a complete graph with N nodes and $N(N-1)/2$ links ($S_{ij} = 0, \pm 1$), obeying the following statistics:

$$\begin{aligned}\langle S_{ij}\rangle_{\mathrm{CC}}^{\mathrm{ER}\langle k\rangle} &= \frac{\langle k\rangle(2q-1)}{N-1} \propto \frac{1}{N},\\ (\Delta S_{ij})_{\mathrm{CC}}^{\mathrm{ER}\langle k\rangle} &= \frac{2\langle k\rangle}{N-1} - \frac{\langle k\rangle^2(2q-1)^2}{(N-1)^2} \propto \frac{1}{N}.\end{aligned} \tag{14}$$

The system has a relevant disorder in the thermodynamic limit and behaves like a genuine spin glass. One expects thus that the thermodynamic limit is already complex. Monte Carlo optimization techniques can offer some hints for the behavior of the r order parameter as a function of q.

There are also some special cases where the r order parameter can be exactly computed. An example for this is the case of tree-like graphs. For such graphs the clusters are simple, and one can obtain the clusters by simply removing the -1 bonds. It is known that at the critical dilution (before falling apart in non-connected sub-graphs) the Erdős–Rényi networks become tree-like, and the average degree of a node is constant in the thermodynamic limit. It is also trivial that for $q = 1$ we have $r = 1$. Now, one can realize that if any finite fraction of the bonds will become negative (i.e., $q < 1$) the order parameter will become less than 1 ($r < 1$). This suggests that the $r(q)$ curves will not have a step-like form for this critically diluted graphs, and contrary to the picture known for globally connected networks the $r = 1$ order parameter is reached only in the $q \to 1$ limit.

3.3. Scale-free networks with fixed average degree

These networks are sometimes known as Barabási–Albert networks [18, 19]. Their degree distribution follows power law, and hence their name as scale-free networks. As discussed in several recent works, these networks are characteristic for many real biological and social systems. In the thermodynamic limit scale-free networks are infinitely diluted complete graphs since their average number of links per node is finite. We expect thus that their statistical properties regarding the CC problem could be again different from those expected for simple Erdős–Rényi networks. There are several methods by which one can generate such networks. The most well-known one is based on a continuous growth governed by preferential attachment [26]. The linear preferential attachment leads to scale-free networks with a power-law exponent -3 for the degree distribution and a fixed average degree, $\langle k\rangle$. For the small size networks (up to several hundreds of nodes) that can be studied by our optimization techniques the network generation method based on preferential attachment is not suitable since the degree distribution will have large deviations from the expected power law. In such cases the less popular configuration model [27, 28] is used to generate the desired

graphs. The configuration model can generate links between a set of nodes, leading to the $n(k)$ degree distribution with a fixed form:

$$n(k) = \alpha k^{-\gamma}. \tag{15}$$

Parameter γ is the power-law exponent ($\gamma \in (0, 2)$) and parameter α governs the number of nodes, N, in the graph. In order to get connected graphs easier it is desirable to have $\gamma \in (0, 2)$. For $\gamma > 2$ most of the generated networks will be disconnected graphs. The advantage of the configuration model is that it generates the desired degree distribution already for relatively small networks.

From the viewpoint of the CC problem we are interested to generate networks with fixed average degree, in order to complete the results obtained in the case of simple Erdős–Rényi type random graphs. In order to get the same $\langle k \rangle$ average degree for different system sizes, both the α and γ parameter values have to be continuously adjusted. The main steps of this graph generation method are the following:

(i) The α and γ parameters are fixed,
(ii) The $n(k)$ ($k = 1, 2, \ldots, k_{\max}$) values are calculated after (15) by truncating them to integer values. The last k value for which $n(k) \geqslant 1$, will be $k_{\max}$. All other $n(k)$ values will be considered as 0.
(iii) The number of nodes, N, and the total number of links, W, in the system are calculated:

$$N = \sum_{k=1}^{k_{\max}} n(k), \qquad W = \frac{1}{2} \sum_{k=1}^{k_{\max}} n(k)k. \tag{16}$$

(iv) The α and γ parameters are changed so that the total number of nodes and the average degree becomes the desired one. This is done by a small exhaustive searching algorithm.
(v) For each $k = 1, 2, \ldots, k_{\max}$ value we assign $n(k)$ nodes that will be the starting point of k links. The endpoints of the links are not yet specified. The number of unspecified end-points is memorized.
(vi) If W is odd, we neglect the link of a node with degree 1, since the total number of such 'floating' links should be even!
(vii) Finally, we connect the nodes by connecting the links with free end points following the procedure from below.
  (a) Two links with free endpoints are randomly selected. If the nodes to which these belong are already connected, another pair is randomly generated.
  (b) We connect the nodes to which the two selected half-links belong and the number of links with free endpoints is decreased by 2.
  (c) continue from step (vii)a until all free end-point links are exhausted.
(viii) Check if the obtained net is a connected one. If not, start again the whole procedure.

There are of course other more sophisticated algorithms which would further improve the scale-free network generation method. One possibility would be to redistribute the fractional parts of the calculated $n(k)$ values, improving the scale-free nature of the networks. For the sake of computational simplicity, in the present study we will use only the above presented simple configuration model.

For scale-free networks with fixed average degree, one would expect in the thermodynamic limit a non-vanishing disorder from the viewpoint of the CC problem. Equation (14) obtained for randomly diluted complete graphs with a fixed average degree holds in this case also. A nontrivial shape for the $r(q)$ curve is expected, and numerical optimization techniques will be used to study finite graphs.

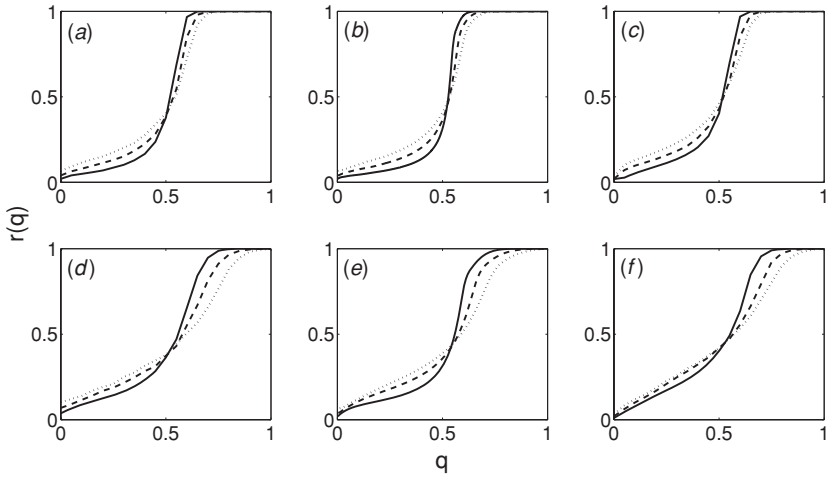

Figure 3. Numerical optimization results for the $r(q)$ curves on randomly diluted networks. For the graphs in the first row (figures 3(a)–(c)) the density of bonds is $p = 0.95$, in the second row (figures 3(d)–(f)) this density is $p = 0.2$. The curves in the first column are generated by the simulated annealing approximation, the curves from the second column by a MC renormalization method, and the curves from the third column are obtained with the MD optimization technique. Different network sizes are considered: $N = 30$ (dotted line), $N = 50$ (dashed line) and $N = 100$ (continuous line).

4. Numerical results

Three optimization methods (briefly discussed in section 2) were used to study the CC problem on networks: simulated annealing, stochastic renormalization and a molecular dynamics approach. Results obtained by these methods are in agreement with each other and support the previously discussed analytical prediction for the thermodynamic limit. Due to the fact that these methods are computationally time consuming and a complicated average has to be done for computing the r order parameter (7), only graphs with relatively modest sizes could be considered. Simulated annealing is the most time-consuming one, and with this method only networks up to 100 nodes could be studied. Larger networks (up to 500 nodes) were studied by the stochastic renormalization and the MD approach. In general, the previously discussed optimization methods were applied on many networks with widely different sizes. In the following however, results for only three different network sizes will be presented, otherwise the graphs would look overcrowded.

In figure 3 we present results for the $r(q)$ curves on randomly diluted networks considering two different dilution rates. Different rows are for different dilutions. In the first row (figures 3(a)–(c)) $p = 0.95$, while in the second row (figures 3(d)–(f)) $p = 0.2$. The graphs in different columns correspond to the different optimization methods. In the first column (figures 3(a)and (d)) simulated annealing results, in the second column (figures 3(b)and (e)) stochastic renormalization results and in the third column MD optimization results (figures 3(c) and (f)) are presented. For each method and dilution several different system sizes ($N = 100$, 50 and 30) were considered. The curves for the MC renormalization technique is smoother due to the fact that this method is faster and consecutively a much smaller step for the variation of q was considered.

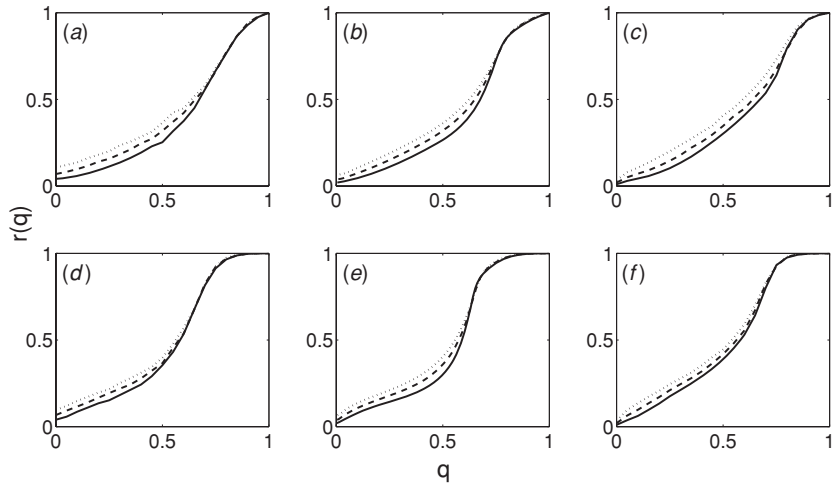

Figure 4. Numerical optimization results for the $r(q)$ curves on randomly diluted networks. For the curves in the first row (figures 4(a)–(c)) the average degree of the nodes is $\langle k\rangle = 4$, for the curves in the second row (figures 4(d)–(f)) the average degree is $\langle k\rangle = 10$. The curves in the first column are generated by the simulated annealing approximation; the curves from the second column by a MC renormalization method and the graphs from the third column are obtained with the MD optimization technique. Results for different network sizes are plotted: $N = 30$ (dotted line), $N = 50$ (dashed line) and $N = 100$ (continuous line).

In the case of the p density of bonds being kept constant, the $r(q)$ curves for increasing system sizes (figures 3(a)–(f)) suggest that in the thermodynamic limit ($N \to \infty$) the same phase-transition is expected as in the globally coupled graphs: $r = 0$ if $q < 0.5$ and $r = 1$ for $q > 0.5$. This is the result suggested in section 3.1 by the simple scaling argument (13). The trend of the curves in figure 3 is pretty similar to those observed in [15] for globally connected systems: as the size of the system increases, the inflection point converges toward $q = 0.5$. This trend is more evident for $p = 0.95$. For the strong dilution case ($p = 0.2$) much bigger system sizes are necessary to have the inflection point close to $q = 0.5$.

Results for finite, randomly diluted networks can be viewed in another perspective, leading to a completely different picture in the thermodynamic limit. In figure 4 we plot the $r(q)$ curves for randomly diluted networks with fixed average degree $\langle k\rangle$. This means that for increasing system size the density of bonds, p, decreases, and in the thermodynamic limit $p \to 0$. Figures 4(a)–(c) are for $\langle k\rangle = 4$ and figures 4(d)–(f) are for $\langle k\rangle = 10$. The graphs in the first column (figures 4(a) and (d)) were obtained by simulated annealing, the graphs in the second column (figures 4(b) and (e)) were obtained by an MC renormalization technique and the graphs in the third column (figures 4(c) and (f)) were generated by the MD optimization technique. For each case systems with three different sizes are presented: $N = 100, 50$ and 30.

For increasing system sizes the $r(q)$ curves exhibit a completely different trend. As has been discussed in section 3.1, in such a case one would expect in the thermodynamic limit a complex spin-glass-type transition, since (14) suggests that the disorder remains relevant. Numerical results confirm this picture. The location of the q_c critical point clearly shifts toward $q_c > 0.5$ values and seemingly the critical point depends on the $\langle k\rangle$ value. Moreover, the curves in figure 4 suggest a completely different scaling trend than the curves in figure 3. As one would naturally expect, for smaller $\langle k\rangle$ values (more diluted systems) the difference

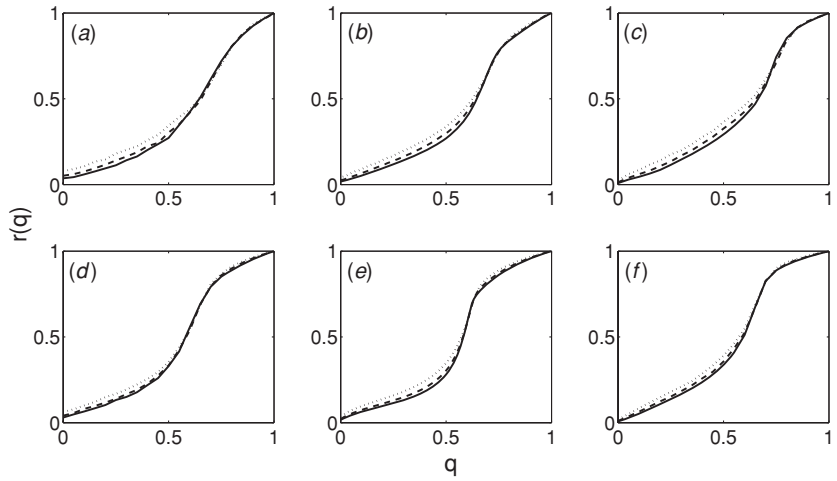

Figure 5. Numerical optimization results for the $r(q)$ curves on scale-free networks. For the graphs in the first row (figures 4(a)–(c)) the average degree of the nodes is $\langle k\rangle = 4$ and for the graphs in the second row (figures 4(d)–(f)) the average degree is $\langle k\rangle = 10$. The curves in the first column are generated by the simulated annealing approximation; the curves from the second column by a MC renormalization method and the curves from the third column are obtained with the MD optimization technique. Different network sizes are considered, in the first row $N = 44$ (dotted line), $N = 75$ (dashed line) and $N = 118$ (continuous line); in the second row: $N = 57$ (dotted line), $N = 99$ (dashed line) and $N = 140$ (continuous line).

Table 1. The parameters of the configuration model for different N and $\langle k\rangle$ values.

N	$\langle k\rangle$	α	γ
118	4	46.0625	1.3
75	4	27.9383	1.2
44	4	13.5991	1.0
140	10	27.9383	0.9
99	10	17.2878	0.8
57	10	7.6141	0.6

from the globally coupled case is more obvious. The trend expected in the thermodynamic limit is obviously different from the trivial step-like behavior at $q_c = 0.5$. Finite-size scaling for $\langle k\rangle = 4$ suggests that in the $q \to 1$ limit the curves are nicely overlapping and suggests that for $q < 1$ one obtains $r < 1$, as it would be expected in the critical dilution limit.

Finally, numerical results for the $r(q)$ curves on scale-free networks were obtained. The configuration model was used to generate scale-free networks of various sizes and with various degree-distribution exponents, γ. In order to have networks with different sizes and fixed average degree one had to continuously adjust the γ value. Networks with $N = 140, 118, 99, 75, 57$ and 44 nodes and with two different $\langle k\rangle$ values were considered: $\langle k\rangle = 4$ and $\langle k\rangle = 10$ (similarly to the case of randomly diluted networks). In table 1 the γ and α values used in generating the studied networks is given.

Results of the numerical optimizations are plotted in figure 5. In the first row (figures 5(a)–(c)) results for $\langle k\rangle = 4$, and in the second row (figures 5(d)–(f)) the results for $\langle k\rangle = 10$ are presented. The curves in the first column (figures 5(a) and (d)) present

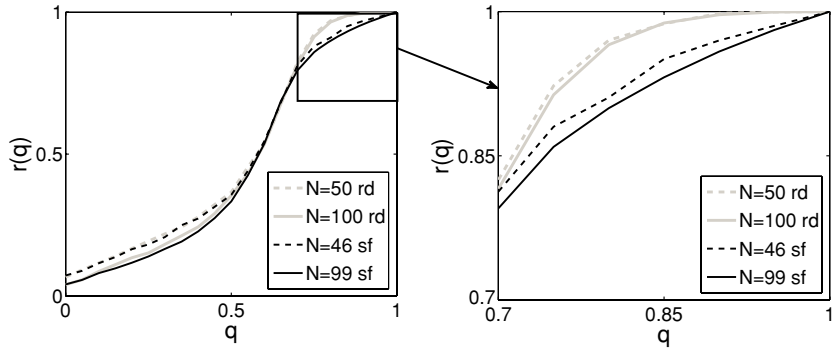

Figure 6. Simulated annealing results for the $r(q)$ curves on randomly diluted and scale-free networks with fixed $\langle k\rangle = 10$ average degree. For each network class two different system sizes are considered as illustrated in the legend. The figure on the right is the magnification of the region inside the box.

results obtained by simulated annealing, the second column (figures 5(b) and (e)) shows the results obtained by MC renormalization and the third column (figures 5(c) and (f)) is for the results obtained by the MD optimization technique. For each method and $\langle k\rangle$ value several system sizes are considered as illustrated in table 1.

The trend of the $r(q)$ curves in figure 5 is similar to those obtained for randomly diluted networks with fixed $\langle k\rangle$ value (figure 4) and suggest a similar complex behavior in the thermodynamic limit. Although the shape of the $r(q)$ curves in figure 4 and figure 5 are similar to each other, plotting them together reveals important differences generated by their specific topology. This is done in figure 6, where the results obtained by simulated annealing are compared for randomly diluted and scale-free networks with $\langle k\rangle = 10$ average degree.

5. Conclusions

The CC problem on randomly diluted and scale-free networks was studied. A simple analytical argument based on the scaling properties of the average value and variance of the links strength (equations (13) and (14)) suggests that in the thermodynamic limit a different complexity class is expected for networks where the bond density, p, is finite and for infinitely diluted graphs ($p = 0$), where the average degree, $\langle k\rangle$, is fixed. For networks with fixed bond density the disorder is irrelevant in the thermodynamic limit and in this limit the same trivial solution is valid as in the case of globally coupled graphs [15, 17]: for the $q < 0.5$ probability of positive links all nodes have to be in separate clusters leading to $r = 0$, while for $q > 0.5$ all nodes have to be in the same cluster ($r = 1$). Contrary to this simple result the problem becomes complex when the average links per node are fixed, meaning in the thermodynamic limit an infinitely strong dilution of the complete graph. In such cases the disorder in the system remains relevant and optimization is a complex NP hard task even in the thermodynamic limit. This simple analytical prediction is verified in the present study by three different numerical optimization approaches on finite networks. The shape of the $r(q)$ curves in the CC problem was numerically studied by simulated annealing, MC renormalization and an MD optimization approach. The results obtained by these methods are all in good agreement with each other and support the predictions of scaling arguments for the thermodynamic limit. This suggests

that the CC problem becomes a complex one even for large networks when the average degree is kept constant. Since scale-free networks are in such a category [18, 19], we expect that the solution of the CC problem in practically interesting social and biological systems remains a complex spin-glass-type optimization problem even for very large networks.

Acknowledgments

Work supported by a Romanian PNCD2/ElBioArch research grant and a Marie Curie International Re-integration Grant within the 6th European Community Framework Program. Gy. Cseh acknowledges financial support of the EU AMPOSDRU PhD program.

References

[1] Bansal N, Blum A and Chawla S 2004 *Mach. Learn.* **56** 89
[2] Charikar M, Guruswami V and Wirth A 2005 *J. Comput. Syst. Sci.* **71** 360
[3] Giotis I and Guruswami V 2006 *Theor. Comput.* **2** 249
[4] Sherrington D and Kirckpatrick S 1975 *Phys. Rev. Lett.* **35** 1792
[5] Cohen W and Richman J 2002 *Proc. 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (New York)* p 475
[6] Kearns M J, Schapire R E and Sellie L M 1994 *Mach. Learn.* **17** 115
[7] Axelrod R and Bennett S 1993 *Br. J. Political Sci.* **23** 211
[8] Galam S 1996 *Physica* A **230** 174
[9] Galam S 1997 *Physica* A **238** 66
[10] Elderfield D and Sherrington D 1983 *J. Phys. C: Solid State Phys.* **16** L497
[11] Erzan A and Lage E J S 1983 *J. Phys. C: Solid State Phys.* **16** L555
[12] Gross D J, Kanter I and Sompolinsky H 1985 *Phys. Rev. Lett.* **55** 304
[13] Kirkpatrick T R and Thirumalai D 1988 *Phys. Rev.* B **37** 5342
Thirumalai D and Kirkpatrick T R 1988 *Phys. Rev.* B **38** 4881
[14] Dillmann O, Janke W and Binder K 1998 *J. Stat. Phys.* **92** 57
[15] Neda Z, Florian R, Ravasz M, Libal A and Gyorgyi G 2006 *Physica* A **362** 357
[16] Garey M R and Johnson D S 1979 *Computers and Intractability. A Guide to the Theory of NP-Completeness* (New York: W. H. Freeman and Company)
[17] Sumi R and Neda Z 2008 *Int. J. Mod. Phys.* C **19** 1349
[18] Albert R and Barabasi A-L 2002 *Rev. Mod. Phys.* **74** 47
[19] Barabasi A-L and Bonabeau E 2003 *Sci. Am.* **288** 60
[20] Barabasi A L *et al* 2002 *Physica* A **311** 590
[21] Kirckpatrick S, Gelatt C D and Vecchi M P 1983 *Science* **220** 671
[22] Rapaport C 1995 *The Art of Molecular Dynamics Simulation* (Cambridge: Cambridge University Press)
[23] Erdos P and Renyi A 1959 *Publicationes Mathematicae* **6** 290
[24] Erdos P and Renyi A 1961 *Acta Math. Sci. Hung.* **12** 261
[25] Bolobas B *Random Graphs* (Academic, London, 1985)
[26] Barabasi A-L, Albert R and Jeong H 1999 *Physica* A **272** 173–87
[27] Bekessy A, Bekessy P and Komlos J 1972 *Stud. Sci. Math. Hungar.* **7** 343
[28] Aiello W, Chung F and Lu L 2000 *Proc. 32nd Annual ACM Symp. on Theory of Computing, Association of Computing Machinery (New York)* p 171