Partitioning and modularity of graphs with arbitrary degree distribution

Jörg Reichardt[*] and Stefan Bornholdt

Institute for Theoretical Physics, University of Bremen, D-28359 Bremen, Germany
(Received 12 June 2006; revised manuscript received 6 April 2007; published 10 July 2007)

We solve the graph bipartitioning problem in dense graphs with arbitrary degree distribution using the replica method. We find the cut size to scale universally with $\langle \sqrt{k} \rangle$. In contrast, earlier results studying the problem in graphs with a Poissonian degree distribution had found a scaling with $\sqrt{\langle k \rangle}$ [Fu and Anderson, J. Phys. A **19**, 1605 (1986)]. Our results also generalize to the problem of q partitioning. They can be used to find the expected modularity Q [Newman and Girvan, Phys. Rev. E **69**, 026113 (2004)] of random graphs and allow for the assessment of the statistical significance of the output of community detection algorithms.

Given a graph or network $\mathcal{G}(N,M)$ of N nodes and M edges, the problem of graph partitioning is finding a partition of the nodes into q equal sized parts, such that the number of edges connecting different parts, the cut size, is minimal. The solution of this problem has many important practical applications in multiprocessor scheduling for parallel computing, very large-scale integrated chip design (VLSI), and data mining [1].

Due to the problem being NP-complete [2], i.e., no algorithm is known which is guaranteed to find an optimal solution in a number of steps that grows only polynomially with the size of the graph, only heuristics exist to approximate a solution. Furthermore, one cannot determine in polynomial time whether a solution found is indeed optimal. The only way to assess the quality of a heuristic algorithm or a particular result is to compare with rigorous bounds on the cut size.

In the past, such bounds have been established for Erdős-Rényi (ER) random graphs [3,4] exhibiting Poissonian degree distribution [5–10] or Bethe lattices with fixed valence [11–15] only. All of these bounds express the expected cut size in terms of the mean degree of the network. However, this is inappropriate for a large number of problem instances which show fat-tailed degree distributions [16,17], and we will show that the traditional bounds are inappropriate in these cases.

Recently, the problem of graph partitioning has received considerable interest again in a more relaxed version, known as community detection or graph clustering which also has a broad range of applications for data analysis in the life sciences [18,19] to social science [20]. Given a graph, the problem is now finding a partition into q arbitrarily sized communities or clusters which are densely connected internally, but only sparsely connected among each other [21,22]. The hypothesis is that, in many real world networks, nodes of differing types exist, and that nodes of the same type have a higher probability of being linked than nodes of different types [23]. Community detection thus aims at discovering precisely these latent classes of nodes.

A widely used quality measure for the partition into communities is the so-called "modularity" Q defined by Newman and Girvan [24]. For a partition into q communities of arbitrary size it is defined as $Q_q = \sum_{s=1}^{q}(e_{ss}-a_s^2)$, with $a_s=\sum_{r=1}^{q}e_{rs}$. Here, e_{rs} is the fraction of edges falling between nodes in groups r and s. Obviously, the larger the distortion between internal edges e_{ss} and expected internal edges a_s^2, i.e., the more modular a network is, the higher Q. By design, $Q < 1$.

In the case of q equal sized parts, we have a relation between the cut size $\mathcal{C}_q$ of the partition and its modularity:

$$\mathcal{C}_q = M\left(\frac{q-1}{q} - Q_q\right). \qquad (1)$$

We see that Q_q measures the improvement of the partition over a random assignment of nodes into parts.

One problem of community detection and all clustering techniques in general is that it is not clear *a priori* whether latent classes of nodes in a particular network do exist. Numerical experiments have shown that random networks can be partitioned in a way which has a very high value of Q, but this of course does not imply that latent classes of nodes in the network [25,26] in fact exist. One way to escape this "deception of randomness" [27] is by comparing the value of Q obtained by clustering the empirical data to an expectation value for Q for an appropriate random null model.

The partition of a purely random graph with maximum modularity is an equipartition [26]. Therefore, by solving the problem of graph partitioning for graphs with arbitrary degree distribution, we can also give expectation values for the maximum attainable modularity in purely random graphs, and vice versa. These expectation values can be used in the assessment of statistical significance of empirical data analysis.

The problem of maximizing the modularity is equivalent to finding the ground state of a spin glass [26,28,29]. The corresponding Hamiltonian $\mathcal{H}$ for $q=2$ relates to the modularity Q_2 at $\gamma=1$ via

$$\langle k \rangle N Q_2 = -\mathcal{H} = \sum_{i<j}(A_{ij}-\gamma p_{ij})S_i S_j, \qquad (2)$$

where $p_{ij}=k_i k_j/\langle k \rangle N$ with $\langle k \rangle$ as the average degree in the network and A_{ij} denoting the adjacency matrix of the graph such that $\Sigma_j A_{ij}=k_i$, the degree of node i. Further, we use Ising spins $S_i=\pm 1$. Following Ref. [30], p_{ij} may be interpreted as the connection probability in the network due to $\Sigma_{i<j}p_{ij}=M=\Sigma_{i<j}A_{ij}$, even though it is not strictly ensured

‾‾‾‾‾‾‾‾‾‾
*Present address: Institute for Theoretical Physics, University of Würzburg, Am Hubland, 97074 Würzburg, Germany.

that $p_{ij} < 1$ for all i, j. The ground state energy of (2) can be written as

$$E_g = -M + 2C_2 + \frac{\gamma}{4M}\left(\sum_{i\in+} k_i - \sum_{i\in-} k_i\right)^2 - \frac{\gamma\langle k^2\rangle}{2\langle k\rangle}. \quad (3)$$

Assuming equal sized parts with the same degree distribution and $\langle k^2\rangle$ being finite, the last two terms in (3) vanish and we see that (1) holds. Due to the existence of $\langle k^2\rangle$, we can also assume the form of p_{ij} to be a good approximation for the true probability distribution that parametrizes A_{ij}.

The Hamiltonian (2) represents an infinite-range spin glass with couplings $J_{ij} = (A_{ij} - \gamma p_{ij})$ between all pairs of nodes. The mean of the coupling matrix is zero by construction for $\gamma = 1$. Interestingly, we are dealing with a system that shows properties of both the Sherrington-Kirkpatrick (SK) [31,32] and the Viana-Bray (VB) type [33]. The distribution of couplings is not symmetric. There are many weak antiferromagnetic couplings J^- between unconnected nodes and few but strong ferromagnetic couplings J^+ between connected nodes:

$$J_{ij}^+ = (1 - \gamma p_{ij}), \quad J_{ij}^- = -\gamma p_{ij}, \quad J_{ij}^+ - J_{ij}^- = J. \quad (4)$$

The parameter γ allows tuning the strength of the antiferromagnetic interactions, and comparing results for different γ enables us to detect hierarchies and overlap in community structures [26,29]. Note that $J = 1$ independent of the choice of the parameter γ. The individual steps of the following development follow closely along the lines of Fu and Anderson (FA) [5], though a treatment along the lines of VB or Monasson [34] is also possible.

We use the replica trick [31,35] in order to evaluate the logarithm of the partition function using $[\ln Z] = \lim_{n\to 0}([Z^n] - 1)/n$. The symbol $[\cdot]$ denotes an expectation value over the ensemble of graphs with a given degree distribution. The probability density function of the coupling matrix J_{ij} is then defined as

$$q_{ij}(J_{ij}) = p_{ij}\delta(J_{ij} - J_{ij}^+) + (1 - p_{ij})\delta(J_{ij} - J_{ij}^-). \quad (5)$$

The p_{ij} here is the same as that in Eqs. (2) and (4). The mean of this distribution is $p(1 - \gamma)$, and we are dealing with a diluted ferromagnet for $\gamma < 1$ and with a diluted antiferromagnet $\gamma > 1$. For all $\gamma \geq 1$ we hence expect zero magnetization in the ground state.

With these preliminaries, we can now follow the steps of FA [5] and average the replicated partition function over the ensemble of graphs with given degree distribution:

$$[Z^n] = \int \prod_{i,j} dJ_{ij}q_{ij}(J_{ij})\text{Tr}_n \exp\left(\beta\sum_\alpha^n \sum_{i<j} J_{ij}S_i^\alpha S_j^\alpha\right). \quad (6)$$

Each of the integrals can be performed separately, and after some manipulation we arrive at

$$[Z^n] = \text{Tr}_n \exp\left[\sum_{i<j} \beta(J_{ij}^- + p_{ij}J)\sum_\alpha S_i^\alpha S_j^\alpha \right.$$
$$\left. + \frac{(\beta J)^2}{2}p_{ij}(1 - p_{ij})\left(\sum_\alpha^n S_i^\alpha S_j^\alpha\right)^2\right], \quad (7)$$

which is still in complete correspondence to Ref. [5]. Now we deviate from FA's development. Instead of a constant $p_{ij} = p$, we use the factorized form introduced earlier. Neglecting terms involving p_{ij}^2 over those involving p_{ij}, dropping all terms which vanish for large N or for $n \to 0$, and rearranging the sums we arrive at

$$[Z^n] = \exp\left(\frac{(\beta J)^2 n\langle k\rangle N}{4}\right)\text{Tr}_n \exp\left[\frac{\beta(J - \gamma)}{\langle k\rangle N}\sum_\alpha \underbrace{\left(\sum_i k_i S_i^\alpha\right)^2}_{m_\alpha^*}\right.$$
$$\left. + \frac{(\beta J)^2}{2\langle k\rangle N}\sum_{\alpha<\beta} \underbrace{\left(\sum_i k_i S_i^\alpha S_i^\beta\right)^2}_{q_{\alpha\beta}^*}\right]. \quad (8)$$

Note that we would have arrived at the same expression by using a Gaussian coupling distribution for $q_{ij}(J_{ij})$ with mean $(J - \gamma)/\langle k\rangle N$ and variance $\sigma_{ij}^2 = k_i k_j/\langle k\rangle N$. At this point, we have a form that is in principle equivalent to the SK model except for the k_i in the sums over spins. We stress again that setting $\gamma = 1$ leads to vanishing magnetization and thus corresponds to a bipartition. The further development now proceeds along the lines of SK and we arrive at one single-site problem for each degree.

Finally, we can derive the replica symmetric equations of state similar to Ref. [36]:

$$m^* = \frac{1}{\langle k\rangle}\sum_k p(k)k \int Dz \tanh[\beta\tilde{H}_k(z)], \quad (9)$$

$$q^* = \frac{1}{\langle k\rangle}\sum_k p(k)k \int Dz \tanh^2[\beta\tilde{H}_k(z)]. \quad (10)$$

The Gaussian measure $Dz = \exp(-z^2/2)/\sqrt{2\pi}$ indicates that the local fields for the spins are also Gaussian distributed, but slightly different for every degree. Nodes of the same degree see the same field distribution. The effective Hamiltonian is

$$\tilde{H}_k(z) = J\sqrt{kq^*}z + (J - \gamma)km^*. \quad (11)$$

We can now again follow standard procedures and find the zero-temperature limit of the free energy and hence Q_2:

$$Q_2 = -\frac{1}{\langle k\rangle}\lim_{\beta\to\infty}[f] = U_0 J\langle k^{1/2}\rangle/\langle k\rangle, \quad (12)$$

where U_0 is the absolute value of the ground state energy per spin of the Sherrington-Kirkpatrick model. It is interesting to note the similarity with results of FA, who give $Q_2 = U_0\sqrt{(1-p)/Np}$ from the ground state energy of the bipartitioning problem on Erdős-Rényi graphs [3,4] with link probability $p_{ij} = p$. For graphs with fixed degree, the two formu-
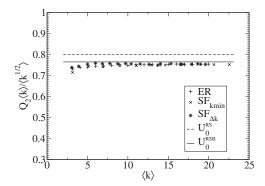
FIG. 1. Numerical experiments maximizing the modularity of bi-partions in random graphs of different topologies [Erdős-Rényi random graphs and two forms of a scale-free (SF) degree distribution; see text for details]. Scaling Q in units of $\langle \sqrt{k} \rangle / \langle k \rangle$ collapses the data points onto one universal curve as expected from (12). The dashed and solid lines correspond to U_0 in the replica symmetric (RS) and replica symmetry breaking (RSB) cases, respectively.

lations of FA and (12) are equivalent. The following numerical experiments will show the adequacy of (12) for a wide range of degree distributions.

First, ER networks of $N = 10\,000$ nodes and average degree between 3 and 20 were created. The second ensemble of networks is that of scale-free (SF) networks with $N = 10\,000$ nodes and a degree distribution of the form $p(k) \propto k^{-\kappa}$ [SF(k_{min})]. The maximum possible degree was set to 1000 and the graphs were composed using the Molloy-Reed algorithm [37]. The second moment of the distribution exists for all $\kappa > 2$. We chose $\kappa = 3$ for our experiments. In order to produce graphs of different average degree, we introduced a minimum degree k_{min}, such that $p(k < k_{min}) = 0$ with $2 \leq k_{min} \leq 12$. Finally, we study a second class of scale-free networks [SF(Δk)]. Introducing a k_{min} may have been too drastic a step, as it excludes all nodes of smaller degree from the network. Therefore, we modify the degree distribution to $p(k) = (k + \Delta k)^{-\kappa}$, using $\kappa = 3$ as before and varying Δk between 2 and 20.

Bipartitions were calculated by minimizing the Hamiltonian (2) using simulated annealing. As expected, the ground state was found to have zero magnetization. Figure 1 shows the result of this experiment. The empirically found maximum modularity for a bipartition in networks from these three classes is rescaled in units of $\langle \sqrt{k} \rangle / \langle k \rangle$ and plotted against $\langle k \rangle$ for better visibility. According to (12) we expect the curves to coincide on a horizontal line at U_0. Indeed, this rescaling collapses the values for different topologies together. The dashed line represents the ground state energy in the replica symmetric (RS) approximation $U_0^{RS} = \sqrt{2/\pi}$ while the solid line corresponds to the full replica symmetry breaking (RSB) ground state energy $U_0^{RSB} = 0.765$ [10,38]. Note that our treatment of the problem allows for the same RSB calculations as the ordinary SK model as all k_i can be drawn out of sums over replica indices. Clearly, we see that (12) gives a good approximation of the modularity and hence cut size for the different topologies. As expected from the mean field nature of our calculation, the estimate is better for denser graphs.
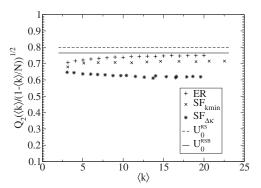

FIG. 2. Same data points as in Fig. 1, but Q in units of $\sqrt{(1-p)/Np}$ as suggested by Fu and Anderson [5]. Clearly, the parametrization by mean degree alone cannot account for the observed cut sizes and modularities in the three different topologies.

In contrast, Fig. 2 shows the same data points as Fig. 1, but this time rescaled in units of $\sqrt{(1-p)/\langle k \rangle}$ as suggested by FA. From the fact that the data points do not collapse onto a single universal curve, we can conclude the inadequacy of this scaling for graphs with a broad degree distribution. Note that (12) also improves over the results of FA in the case of ER graphs.

Our results from the case of bipartitioning generalize in a straightforward way to the case of q partitioning, and we can simply replace the scaling from the formulas by Kanter and Sompolinsky (KS) [6] or Lai and Goldschmidt [7]. The expectation value for the maximum modularity of a random graph with arbitrary degree distribution is then [26]

$$Q = 0.97 \frac{\langle \sqrt{k} \rangle}{\langle k \rangle}. \tag{13}$$

The factor of $U_0 = 0.97$ corresponds to the ground state energy of the Potts glass in the one-step RSB treatment as
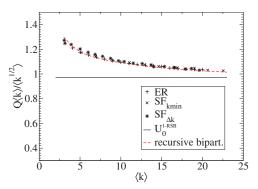

FIG. 3. (Color online) Numerical experiments maximizing the modularity in random graphs of different topologies [Erdős-Rényi (ER) random graphs and two forms of a scale-free (SF) degree distribution; see text for details]. Scaling Q in units of $\langle \sqrt{k} \rangle / \langle k \rangle$ collapses the data points onto one universal curve as expected from (12). The solid lines correspond to ground state energy U_0 of the Potts glass in the one-step replica symmetry breaking approximation due to Kanter and Sompolinsky [6]. The dashed line corresponds to an estimate obtained from a recursive bipartitioning along the lines of Ref. [26] using Eq. (12).

calculated by KS [6,26]. Figure 3 shows the maximum modularity obtained when minimizing the Hamiltonian (2) in the same graphs as used in Fig. 1 but with the number of communities q as a free parameter. Again, we see that plotting Q in units of $\langle\sqrt{k}\rangle/\langle k\rangle$ collapses the data points onto a single universal curve. The approximation of the universal value of U_0, however, is much slower. This was already observed in Ref. [26] and can be explained by the fact that sparser graphs tend to cluster into more modules than predicted by KS. This leads to a higher number of degrees of freedom which then tend to accommodate better for fluctuations in the link structure of sparse graphs which lead to relatively higher modularities or lower cut sizes than expected for denser graphs. Since the rescaling of the modularity has made the data points from all topologies collapse onto a single universal curve, we can employ a recursive estimate using Eq. (12) several times as outlined in Ref. [29]. It is interesting to note that, once we have collapsed the data onto a universal curve, we can use the recursive estimate from the convenient ER graphs to give analytical bounds for other topologies as well.

In conclusion, we have presented different approxima-tions for the cut size of a bipartition and the corresponding modularity in dense graphs with arbitrary degree distributions. Our results show that scale-free networks exhibit larger cut sizes and smaller modularities than ER graphs of the same average degree. Graphs with a modified scale-free degree distribution $p(k)\propto(k+\Delta k)^\kappa$ have larger cut sizes and smaller modularities than those with a minimum degree. This result also gives insight into the stability of networks to targeted attack (link removal) [39]. The estimation of modularity of random graphs presented allows for a convenient comparison of the output of community detection algorithms with appropriate random null models and hence gives a clue to the statistical significance of the findings in real world networks and paves the way to understanding the theoretical limits of community detection [40].

[1] E.-H. Han, G. Karypis, V. Kumar, and B. Mobasher, Data Eng. Bull. **21**, 15 (1998).

[2] M. R. Gareya and D. S. Johnson, *Computers and Intractability* (W. H. Freeman & Co., London, 1979).

[3] P. Erdős and A. Rényi, Publ. Math. Inst. Hung. Acad. Sci. **5**, 17 (1960).

[4] B. Bollobas, *Random Graphs*, 2nd ed. (Cambridge University Press, Cambridge, U.K., 2001).

[5] Y. Fu and P. W. Anderson, J. Phys. A **19**, 1605 (1986).

[6] I. Kanter and H. Sompolinsky, J. Phys. A **20**, L636 (1987).

[7] P.-Y. Lai and Y. Y. Goldschmidt, J. Stat. Phys. **48**, 513 (1987).

[8] W. Liao, Phys. Rev. A **37**, 587 (1988).

[9] W. Liao, Phys. Rev. Lett. **59**, 1625 (1987).

[10] Y. Y. Goldschmidt and C. D. Dominicis, Phys. Rev. B **41**, 2184 (1990).

[11] K. Y. M. Wong and D. Sherrington, J. Phys. A **20**, L793 (1987).

[12] M. Mezard and G. Parisi, Europhys. Lett. **3**, 1067 (1987).

[13] J. R. Banavar, D. Sherrington, and N. Sourlas, J. Phys. A **20**, L1 (1987).

[14] M. J. de Oliveira, J. Stat. Phys. **54**, 477 (1989).

[15] D. Sherrington and K. Y. M. Wong, J. Phys. A **20**, L785 (1987).

[16] M. Steinbach, L. Ertöz, and V. Kumar, *New Vistas in Statistical Physics—Applications in Econo-physics, Bioinformatics, and Pattern Recognition* (Springer-Verlag, Berlin, 2004), Chap. 2.

[17] C.-K. Cheng, A. B. Kahng, and B. Liu, in *SLIP '01: Proceedings of the 2001 International Workshop on System-Level Interconnect Prediction* (ACM Press, New York, 2001), pp. 99–106.

[18] R. Guimera and L. A. N. Amaral, Nature (London) **433**, 895 (2005).

[19] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, Nature (London) **435**, 814 (2005).

[20] A. Arenas, A. Diaz-Guilera, P. Gleiser, and R. Guimera, Eur. Phys. J. B **38**, 373 (2004).

[21] M. E. J. Newman and M. Girvan, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2003).

[22] L. Danon, J. Dutch, A. Arenas, and A. Diaz-Guilera, J. Stat. Mech.: Theory Exp. (2005) P09008.

[23] M. E. J. Newman, Phys. Rev. E **67**, 026126 (2003).

[24] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).

[25] R. Guimera, M. Sales-Pardo, and L. A. N. Amaral, Phys. Rev. E **70**, 025101(R) (2004).

[26] J. Reichardt and S. Bornholdt, Phys. Rev. E **74**, 016110 (2006).

[27] A. Engel and C. V. den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, UK 2001).

[28] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).

[29] J. Reichardt and S. Bornholdt, Phys. Rev. Lett. **93**, 218701 (2004).

[30] F. Chung and L. Lu, Ann. Comb. **6**, 125 (2002).

[31] D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett. **35**, 1792 (1975).

[32] S. Kirkpatrick and D. Sherrington, Phys. Rev. B **17**, 4384 (1978).

[33] L. Viana and A. J. Bray, J. Phys. C **18**, 3037 (1985).

[34] R. Monasson, J. Phys. A **31**, 513 (1998).

[35] S. F. Edwards and P. W. Anderson, J. Phys. F: Met. Phys. **5**, 965 (1975).

[36] D.-H. Kim, G. J. Rodgers, B. Kahng, and D. Kim, Phys. Rev. E **71**, 056115 (2005).

[37] M. Molloy and B. Reed, Random Struct. Algorithms **6**, 161 (1995).

[38] G. Parisi, J. Phys. A **13**, L115 (1980).

[39] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **406**, 378 (2000).

[40] J. Reichardt and S. Bornholdt, Physica D **224**, 20 (2006).