Modularity from fluctuations in random graphs and complex networks

Roger Guimerà, **Marta Sales-Pardo**, and **Luís A. Nunes Amaral**
Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, USA

Abstract

The mechanisms by which modularity emerges in complex networks are not well understood but recent reports have suggested that modularity may arise from evolutionary selection. We show that finding the modularity of a network is analogous to finding the ground-state energy of a spin system. Moreover, we demonstrate that, due to fluctuations, stochastic network models give rise to modular networks. Specifically, we show both numerically and analytically that random graphs and scale-free networks have modularity. We argue that this fact must be taken into consideration to define statistically significant modularity in complex networks.

Statistical, mathematical, and model-based analysis of complex networks have recently uncovered interesting unifying patterns in networks from seemingly unrelated disciplines [1-5]. In spite of these advances, many properties of complex networks remain elusive, a prominent one being modularity [6,7]. For example, it is a matter of common experience that social networks have communities of highly interconnected nodes that are poorly connected to nodes in other communities. Such modular structures have been reported not only in social networks [6-8], but also in biochemical networks [9], food webs [10], and the Internet [11]. It is widely believed that the modular structure of complex networks plays a critical role in their functionality [9]. There is therefore a clear need to develop algorithms to identify modules accurately [6,7,11-13].

More fundamentally, the mechanisms by which modularity emerges in complex networks are not well understood. In biological networks—both biochemical and ecological—researchers have suggested that modularity increases robustness, flexibility, and stability [9,10]. Similarly, in engineered networks, it has been suggested that modularity is effective to achieve adaptability in rapidly changing environments [14]. It may therefore seem that *evolutionary pressures* make networks modular, implying that any successful model of complex networks should take into account external factors that enhance modularity. Recently, however, Solé and Fernàndez have pointed out that models without any external pressure are able to give rise to modular networks [15].

In this paper, we show that Erdös-Rényi (ER) random graphs, in which any pair of nodes is connected with probability p [16], have a high modularity. We show numerically and analytically that this high modularity is due to fluctuations in the establishment of links, which are *magnified* by the large number of ways in which a network can be partitioned into modules. Furthermore, we show that one obtains similar results when considering scale-free networks [2]. We conclude by discussing how these results should be taken into consideration to define statistically significant modularity in complex networks.

Following the first quantitative definition of modularity [7,12], several groups have proposed heuristic algorithms to detect modules in complex networks. For a given partition of the nodes of a network into modules, the modularity $\mathcal{M}$ of this partition is defined as [7]

$$\mathcal{M} \equiv \sum_{s=1}^{r} \left[\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right],$$

(1)

where r is the number of modules, L is the number of links in the network, l_s is the number of links between nodes in module s, and d_s is the sum of the degrees of the nodes in module s. This definition of modularity implies that $\mathcal{M} \leq 1$ and that $\mathcal{M} = 0$ for a *random partition* of the nodes [7]. We define the modularity M of a network as the largest modularity of all possible partitions of the network $M = \max\{\mathcal{M}\}$.

The problem of finding the modularity of a network with S nodes is therefore analogous to the standard statistical mechanics problem of finding the ground-state energy of the Hamiltonian $\mathcal{H} = -L\mathcal{M}$. Specifically, one can map the network into a spin system by defining the variables $s_i \in \{1,2,\dots,S\}$ as the module to which node i belongs and the couplings J_{ij} as being 1 if nodes i and j are connected in the network and 0 otherwise. Then, from Eq. (1), one can demonstrate that

$$\mathcal{H} = - \sum_{i,j}^{S} \frac{J_{ij}}{2} \delta_{s_i s_j} + \sum_{i,j,k,l}^{S} \left[\frac{J_{ij}J_{kl}}{16L} \left(\delta_{s_i s_j s_k s_l} + 2\delta_{s_i s_j s_k} + \delta_{s_i s_k} \right) \right].$$

(2)

This Hamiltonian corresponds to an S-state Potts model with both ferromagnetic and anti-ferromagnetic terms, and two-, three-, and four-spin interactions. Therefore, it seems difficult to apply methods used in problems that are similar but formally simpler, like the graph coloring problem [17]. Rather, we propose here a heuristic estimation of the modularity for a number of interesting graph models, namely low-dimensional regular lattices, ER random graphs [16] and scale-free networks [2].

Low-dimensional regular lattices

Consider a one-dimensional lattice with S nodes, each one connected to its two neighbors [20]. This case is particularly simple because the modules comprise only contiguous nodes and, therefore, the number of between-module links equals the number r of modules. Assuming that all modules have approximately the same size $n = S/r$, the modularity of a partition with r modules is

$$\mathcal{M}_{1D}(S;r) = \frac{S - r}{S} - \frac{1}{r},$$

(3)

where we have used the fact that the number L of links is $L \approx S$. Under these assumptions, the problem of finding the modularity of a regular one-dimensional lattice is reduced to finding the optimal number r^* of modules, that is, the number of modules that yields the maximum modularity. One can show that $r^*(S) = \sqrt{S}$, and the modularity is

$$M_{1D}(S) = 1 - \frac{2}{\sqrt{S}}.$$

(4)

Note that the only assumption in the calculation is that all modules have approximately the same number of nodes. Numerical results confirm that this is a sensible assumption.

One can generalize this result to one-dimensional lattices in which each node is connected to z nodes on the left and z on the right. In this case, the leading contributions to the modularity are

$$M_{1D}(S,z) = 1 - \sqrt{\frac{2(z+1)}{S}}.$$

(5)

Similarly, one can calculate the modularity of d-dimensional cubic lattices in which each node is connected to $2z$ nodes in each one of the d directions, to obtain that [18]

$$M_{dD}(S,z) = 1 - (d+1)\left(\frac{z+1}{2d}\right)^{d/(d+1)}\frac{1}{S^{1/(d+1)}}.$$

(6)

Random graphs

In ER random graphs [16], each pair of nodes is connected with probability p. As for d-dimensional lattices, we assume that the partition of the network with highest modularity consists of r modules with approximately the same number of nodes $n=S/r$, the same number of within-module links k_i, and the same number of links k_o to other modules. In the $S \gg 1$ limit, we can assume that the total number of links is $S^2p/2$ and, therefore, k_i and k_o are related by

$$k_o(S,p;r,k_i) = \frac{S^2p}{r} - 2k_i.$$

(7)

Hence, for $S \gg 1$, the modularity of such a partition is simply

$$\mathcal{M}_{ER}(S,p;r,k_i) = \frac{2rk_i}{S^2p} - \frac{1}{r}.$$

(8)

Under these assumptions, the problem of finding the modularity of a random graph is reduced to finding a partition of the graph with the following properties: (i) The partition consists of r equal modules, each one with k_i within-module links; (ii) the partition *typically exists* in a random graph; and (iii) the partition yields the maximum modularity relative to the other partitions that typically exist.

In a random graph with S nodes and linking probability p, the average number $\mathcal{N}$ of different partitions with r identical modules, each with k_i links, is $\mathcal{N}(S,p;r,k_i)$. A certain partition *typically exists* if $\mathcal{N}(S,p;r,k_i) \geq 1$. Among all the partitions that typically exist, we are interested in the one whose modularity is maximum. In other words, given a certain number r of modules, we want a partition with as many within-module links as possible. Therefore, if one finds a very *common* partition $\mathcal{N}(S,p;r,k_i) \geq 1$, it must be possible to find another partition with the same r and $k_i' > k_i$ that has larger modularity. This new partition will be *rarer* than the former one $\mathcal{N}(S,p;r,k_i') < \mathcal{N}(S,p;r,k_i)$. By iterating this argument, one concludes that the partition we are interested in must satisfy

$$\mathcal{N}(S,p;r,k_i^*(S,p;r)) = 1,$$

(9)

where $k_i^*(S,p;r)$ is the maximum number of within-module links that one can typically find in a partition with r identical modules.

To calculate $\mathcal{N}(S,p;r,k_i)$, we use the following process. First, we calculate the number $\mathcal{N}_1$ of ways in which a module of size $n=S/r$, with k_i within-module links and $k_o(r,k_i)$ external links, can be *separated* from the rest of the graph:

$$\mathcal{N}_1 = \binom{S}{n} P_i(S,p;n,k_i) P_o(S,p;n,k_o),$$

(10)

where

$$P_i(S,p;n,x) = \binom{\frac{n(n-1)}{2}}{x} p^x(1-p)^{[n(n-1)]/2-x},$$

(11)

$$P_o(S,p;n,x) = \binom{n(S-n)}{x} p^x (1-p)^{n(S-n)-x}.$$

(12)

The next step is to separate the second module from the remaining set of $S-n$ nodes. It is important to note that the second module only needs to establish $k_o(1-n/(S-n))$ external links, because the remaining $k_o n/(S-n)$ are already established with the first module. Therefore,

$$\mathcal{N}_2 = \binom{S-n}{n} P_i(S,p;n,k_i) P_o\left(S,p;n,k_o\left(1-\frac{n}{S-n}\right)\right).$$

(13)

Repeating this separation process, one can see that the general term is of the form

$$\mathcal{N}_{t+1} = \binom{S-tn}{n} P_i(S,p;n,k_i) P_o\left(S,p;n,k_o\left(1-\frac{tn}{S-n}\right)\right).$$

(14)

Finally, $\mathcal{N}(S,p;r,k_i)$ is the product of all the individual module separations

$$\mathcal{N}(S,p;r,k_i,k_o(r,k_i)) = \prod_{t=1}^{r} \mathcal{N}_t,$$

(15)

so that Eq. (9) can be solved numerically to obtain $k_i^*(S,p;r)$ using Eqs. (11), (12), (14), and (15).

Once we find $k_i^*(S,p;r)$ for a given value of r, we use Eq. (8) to obtain the modularity. Finally, we select the optimal number of modules $r=r_*(S,p)$ and the modularity $M_{\mathrm{ER}}(S,p)$ of the ER random graph is

$$M_{\mathrm{ER}}(S,p) = \frac{2r^*(S,p)\,k_i^*(S,p;r^*)}{S^2 p} - \frac{1}{r^*(S,p)}.$$

(16)

In Fig. 1(a), we compare the modularity of ER graphs obtained through optimization of Eq. (1) using simulated annealing [19], with the predictions of Eq. (16). We find good agreement in the relevant region of sparse but connected graphs, that is, $2/S,<p\ll1$.

Equation (16) enables us to obtain the modularity of large random graphs, something that would not be possible using simulated annealing because of the computational cost. In Fig. 1(b) we show that for $S\rightarrow\infty$ the modularity only depends on pS

$$M_{\mathrm{ER}}(S\rightarrow\infty,p)\langle(pS)^{-2/3}.$$

(17)

To obtain a closed expression for M_{ER} for any value of S, we note that at the percolation point $pS=2$ the random graph contains essentially no loops, that is, the graph is a tree [16]. In this case, one can find partitions in which the number of between-module links equals the number of modules r as in the simple one-dimensional case, and the modularity is

$$M_{\mathrm{ER}}(S,p=2/S) = M_{1D}(S) = 1 - \frac{2}{\sqrt{S}}.$$

(18)

We propose the simplest *ansatz* that verifies Eqs. (17) and (18) simultaneously

$$M_{\mathrm{ER}}(S,p) = \left(1 - \frac{2}{\sqrt{S}}\right)\left(\frac{2}{pS}\right)^{2/3}.$$

(19)

In Fig. 1(a), we show that Eq. (19) is in good agreement with values obtained using simulated annealing.

Our analytic treatment allows us to explain the origin of the modularity in random graphs. The typical partition of an ER graph into modules of size n is very unlikely to have a number of within-module links k_i larger than the average $pn(n-1)/2$, expected for a random partition of the nodes. However, the number of possible partitions $S!/(n!r)$ is so large that, typically, there exists a partition whose k_i is much larger than the average. For example, for a network with S=200 and p=0.02 one typically finds a partition with r=7 modules and $k_i \approx$ instead of the value $k_i \approx 8$ expected for a random partition.

Remarkably, the modularity of a random graph can be as large as that of a graph with modular structure imposed at the onset [6]. In such a graph, nodes are divided into modules and each pair of nodes is connected with probability p_i if they belong to the same module, and with probability $p_o > p_i$ otherwise. Using the same example as before, the modularity of an ER graph with S=200 and p=0.02 is the same as the modularity of a graph with m=7 modules, $p_i \approx 0.09$, and $p_o \approx 0.004$.

Scale-free networks

So far, we have considered d-dimensional regular lattices and ER random graphs, in which all nodes have essentially the same degree. However, many complex networks display scale-free degree distributions [4], meaning that some nodes have degrees that are orders of magnitude larger than the average. Since the results presented for ER graphs rely on the fact that there are many partitions of the network and implicitly on the fact that nodes are exchangeable, it is worth asking whether "random" scale-free networks also display modularity.

To answer this question, we use the scale-free model proposed in [2]. In the model, the network grows by the addition of new nodes. Each time a new node is added, it establishes m preferential connections to nodes already in the network. In Fig. 2, we show the modularity of scale-free networks as a function of the network size S for different values of m. As before, we find the modularity by optimizing Eq. (1) using simulated annealing. As for ER graphs, the modularity approaches a finite value for large S and decreases with the connectivity m.

We are unable to derive a general expression for the modularity of scale-free networks. However, for m=1 the scale-free network is a tree. Thus,

$$M_{SF}(S,m=1) = M_{1D}(S) = 1 - \frac{2}{\sqrt{S}}.$$

(20)

For larger values of m, we find numerically that, at a fixed network size, the modularity is a linear function of $1/m$. The simplest possible *ansatz* for the modularity that verifies this condition and Eq. (20) simultaneously is

$$M_{SF}(S,m) = \left(a + \frac{1-a}{m}\right)\left(1 - \frac{2}{\sqrt{S}}\right).$$

(21)

As we show in Fig. 2, this approximation works well for a=0.165±0.009.

Conclusions

We have shown that modularity in networks can arise due to a number of mechanisms. We have demonstrated that networks embedded in low-dimensional spaces have high modularity. We have also shown analytically and numerically that, surprisingly, random graphs and scale-free networks have high modularity due to fluctuations in the establishment of links.

Recently, several works have reported the existence of modules in complex networks and suggested that some evolutionary mechanism must enhance modularity. This statement is

based, in the best cases, on the fact that the modularity is *large enough*, and relies implicitly on the assumption that random graphs have low modularity.

Our results enable one to define *statistically significant* modularity in networks. We argue that, just as it is already done for the clustering coefficient and other quantities, the modularity of complex networks must always be compared to the *null case* of a random graph. The analytical expressions we have derived provide a convenient way to carry out such a comparison.

References

1. Watts DJ, Strogatz SH. Nature (London) 1998;393:440. [PubMed: 9623998]

2. Barabási A-L, Albert R. Science 1999;286:509. [PubMed: 10521342]

3. Amaral LAN, Scala A, Barthelémy M, Stanley HE. Proc. Natl. Acad. Sci. U.S.A 2000;97:11149. [PubMed: 11005838]

4. Albert R, Barabási A-L. Rev. Mod. Phys 2002;74:47.

5. Dorogovtsev SN, Mendes JFF. Adv. Phys 2002;51:1079.

6. Girvan M, Newman MEJ. Proc. Natl. Acad. Sci. U.S.A 2002;99:7821. [PubMed: 12060727]

7. Newman MEJ, Girvan M. Phys. Rev. E 2004;69:026113.

8. Guimerà R, Danon L, Diaz-Guilera A, Giralt F, Arenas A. Phys. Rev. E 2003;68:065103(R).Arenas A, Danon L, Díaz-Guilera A, Gleiser PM, Guimerà R. Eur. Phys. J. B 2004;38:373.

9. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. Nature (London) 1999;402:C47. [PubMed: 10591225]Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási A-L. Science 2002;297:1551. [PubMed: 12202830]Holme P, Huss M. Bioinformatics 2003;19:532. [PubMed: 12611809]

10. Pimm SL. Theor Popul. Biol 1979;16:144. [PubMed: 538731]Krause AE, Frank KA, Mason DM, Ulanowicz RE, Taylor WW. Nature (London) 2003;426:282. [PubMed: 14628050]

11. Eriksen KA, Simonsen I, Maslov S, Sneppen K. Phys. Rev. Lett 2003;90:148701. [PubMed: 12731952]

12. Newman MEJ. Phys. Rev. E 2004;69:066133.

13. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D. Proc. Natl. Acad. Sci. U.S.A 2004;101:2658. [PubMed: 14981240]Reichardt J, Bornholdt S. e-print cond-mat/0402349Fortunato S, Latora V, Marchiori M. e-print cond-mat/0402522Capocci A, Servedio VDP, Caldarelli G, Colaiori F. e-print cond-mat/0402499

14. Alon U. Science 2003;301:1866. [PubMed: 14512615]

15. Solé RV, Fernández P. e-print q-bio.GN/0312032

16. Bollobas, B. Random Graphs. 2nd ed.. Cambridge University Press; New York: 2001.

17. Mulet R, Pagnani A, Weigt M, Zecchina R. Phys. Rev. Lett 2002;89:268701. [PubMed: 12484862]van Mourik J, Saad D. Phys. Rev. E 2002;66:056120.

18. Guimerà R, Sales-Pardo M, Amaral LAN. unpublished

19. Kirkpatrick S, Gelatt CD, Vecchi MP. Science 1983;220:671. [PubMed: 17813860]

20. In this case, there are no fluctuations involved in the creation of the network. Rather, modularity arises because neighbors of a node in a low-dimensional lattice are also neighbors of each other, and neither the node nor its neighbors are linked to nodes that are far away in the lattice. As an example, consider the cities and towns in Europe and all the roads between them. This "road network" is two-dimensional and has modules that correspond roughly to the countries. In each of this "sub-modules" people have different customs, food preferences, languages or dialects, etc., that is, communities exist. It is also worth noting that the hypothesis used in the calculations are essentially the same that we use for random graphs and scale-free networks, and that the modularity of one-dimensional regular lattices turns out to be useful in certain limits of ER random graphs and scale-free networks.
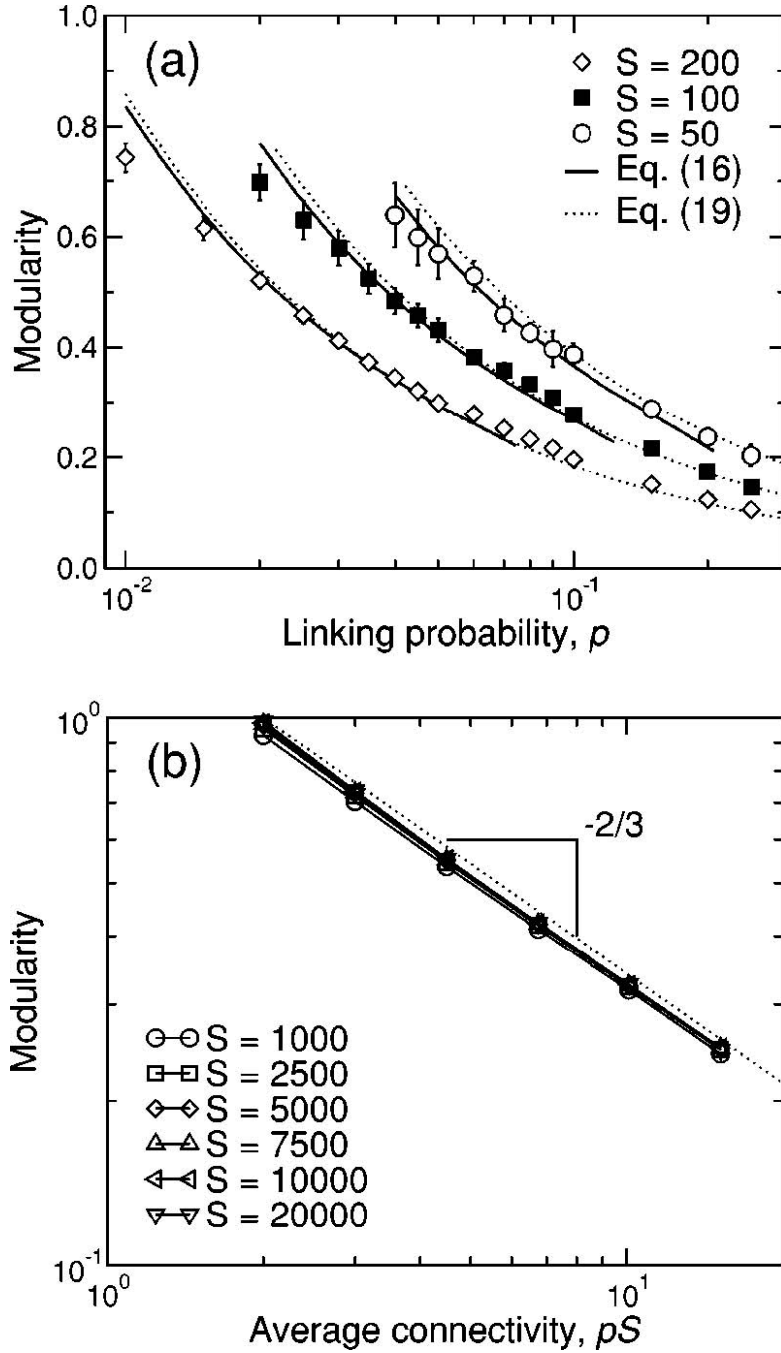
FIG. 1.
Modularity in Erdös-Rényi random graphs. (a) Comparison of numerical results of the modularity as a function of the linking probability, and the predictions of Eqs. (16) and (19). The numerical results are obtained by maximizing the modularity, Eq. (1), using simulated annealing [19]. (b) Modularity as a function of pS for large networks, as predicted by Eq. (16). Both in (a) and (b), numerical problems in the solution of Eq. (9) prevent us from obtaining values of the modularity for larger values of p.
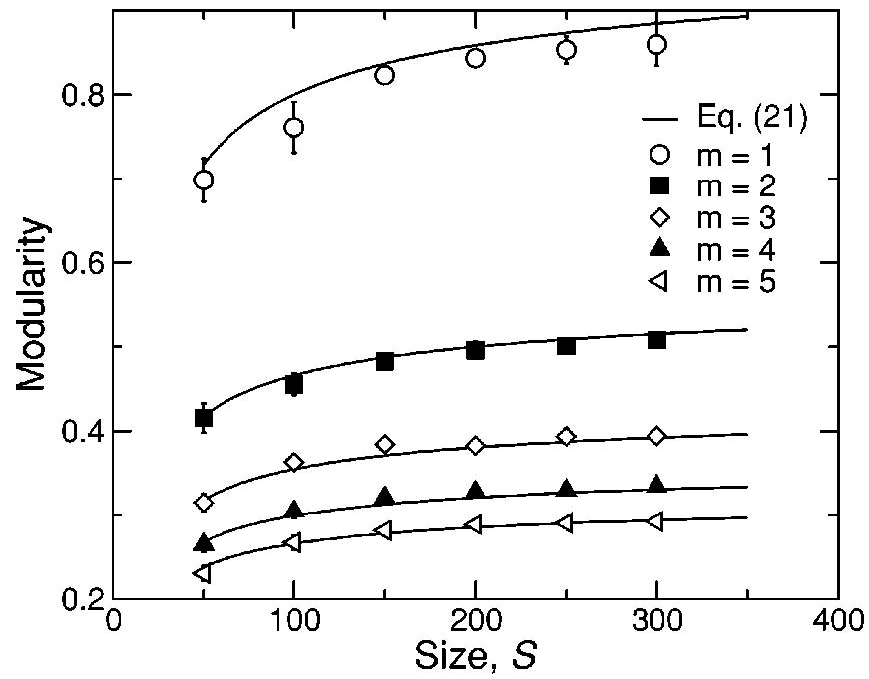
FIG. 2.
Modularity in scale-free networks. Numerical results of the modularity as a function of the network size *S* for different values of *m*. These results are obtained by maximizing the modularity, Eq. (1), with simulated annealing. The lines are the predictions of Eq. (21), with *a*=0.165±0.009 in all the cases.