1

RANDOM GRAPHS: TYPICAL AND RARE PROPERTIES

1.1 Statistical ensembles of random graphs

1.1.1 *Poissonian graphs*

Two popular models of random graphs over N vertices are,

- *Model I (fixed number of edges):*
  Consider the complete graph K_N over N vertices. We define $\mathcal{G}_{N,N_E}$ as the set of graphs obtained by taking only $N_E = cN/2$ among the $\binom{N}{2}$ edges of K_N in all possible different ways. Within Distribution I, a random graph G is a randomly chosen element of $\mathcal{G}_{N,N_E}$ with the flat measure,

$$\mathcal{P}_{\mathcal{I}}(G) = \frac{1}{\binom{\binom{N}{2}}{N_E}} \qquad . \tag{1.1}$$

- *Model II (fixed probability of edge deletion):*
  Another way of generating random graphs from the complete graph is through edge deletion. Start from K_N, and delete every edge with probability $1 - c/N$. This choice ensures that the average number of edges left at the end of the deletion process equals N_E defined in Distribution I. The probability $\mathcal{P}(G)$ of drawing a random graph G depends on the number $N_E(G)$ of its edges, and follows the Binomial law,

$$\mathcal{P}_{\mathcal{II}}(G) = \left(\frac{c}{N}\right)^{N_E(G)} \left(\mathit{1} - \frac{c}{N}\right)^{\binom{N}{2} - N_E(G)} \qquad . \tag{1.2}$$

In the large N limit, both distributions of random graphs share common properties which depend on the control parameter c. We shall make explicit the distribution considered when necessary.[1]

Fig. 1.1 shows examples of graphs obtained from Distribution I and for various values of c. Notice the qualitative change of structure of graphs when the connectivity c varies from low values (graphs are mostly made of small isolated trees) to higher ones (a large part of vertices are now connected together). This

[1] Distribution I is easier to implement on a computer. Drawing a graph amounts to chose N_E distinct pairs of vertices among the $N(N-1)/2$ possible ones, a task which can be carried out in $O(N_E)$ $(= O(N)$ for finite $c)$ steps. On the contrary, within Distribution II, all $O(N^2)$ edges of K_N have to been looked at in the course of the deletion process. However, Distribution II does not induce any correlation between different edges and is therefore more convenient from an analytical point of view.

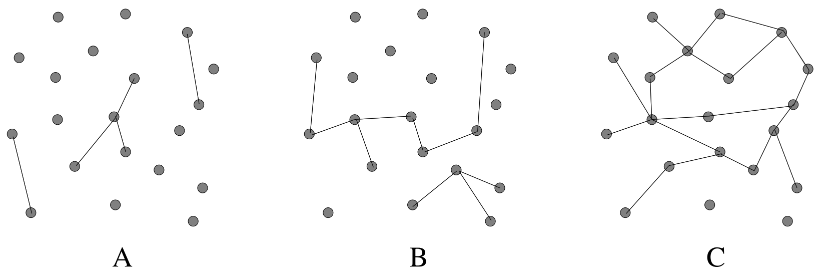

FIG. 1.1. Examples of random graphs generated at fixed number N_E of edges (Distribution I). All graph include $N = 20$ vertices (grey dots). The average degrees of valency, $c = 2N_E/N$, equal $c = 0.5$ (**A**), $c = 1$ (**B**), and $c = 2$ (**C**). The labels of the vertices have been permuted to obtain planar graphs, *i.e.* avoid crossing of edges.

change is known as the percolation transition in physics, or the appearance of a giant component in mathematics literature.

Before reviewing some of the aspects of the percolation transition rigorously established by mathematicians, let us mention an important fact on the valency of vertices. As a result of the randomness of the graph generation process, each node share edges with a variable number of neighboring vertices. Both random graph ensembles are called Poissonian since, in the large N limit, the degree v of a vertex, *i.e.* the number of its neighbors, is a random variable obeying a Poisson law with parameter c,

$$\rho(v) = \lim_{N\to\infty} \binom{N}{v} \left(\frac{c}{N}\right)^v \left(1-\frac{c}{N}\right)^{(N-1)-v} = e^{-c}\,\frac{c^v}{v!} \qquad . \tag{1.3}$$

In particular, $\rho(0) = e^{-c}$ is the fraction of isolated vertices. Control parameter c may be thus seen as the average degree of nodes. This is what is meant in the following when referring to c as the connectivity of the graph.

1.1.2 *Other ensembles*

Random graphs exist for while the distribution $\rho(v)$ of the degrees of vertices is not Poissonian. For instance the K-regular random graph ensemble gives uniform weight to all graphs where every vertex has degree K exactly, zero weight to the other graphs. Particular attention has recently been brought to the case of algebraically decreasing laws *i.e.* $\rho(v) \propto v^{-\tau}$ at large v; this power law behaviour is supposed to reflect the properties of various graphs ranging from technological applications to biological networks. Let us introduce the generating function of the degree probabilities,

$$G_0(x) = \sum_{v\geq 0} \rho(v)\; x^v \qquad . \tag{1.4}$$

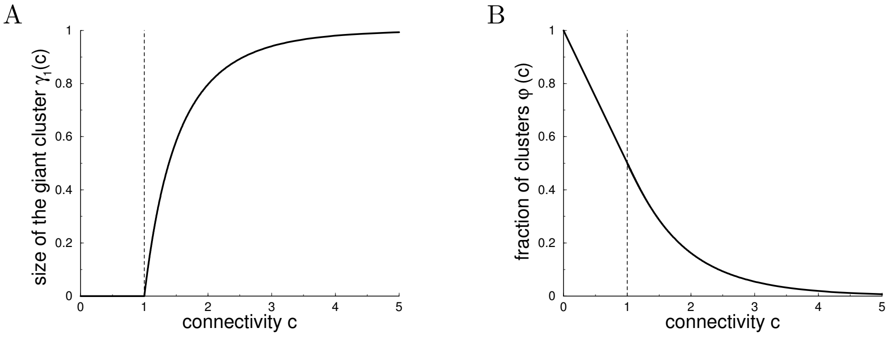

FIG. 1.2. The percolation transition in the random graph. Fraction $\gamma_1(c)$ of vertices in the largest component (**A**), and number $\varphi^*(c)$ of clusters per vertex (**B**) a function of the average connectivity c of the random graph. The vertical dot–dashed line $c = 1$ indicates the location of the percolation threshold.

Clearly, $G_0(1) = 1$ since ρ is normalised, and G_0 is an absolulетly convergent series in x over $[0;1[$. Furthermore, we assume that the average degree is well defined *i.e.* the derivative of G_0 in $x = 1$ is finite, and denote its value by v_0. We define the probability $\rho_1(v')$ that a node connected to a randomly chosen edge in the graph has itself v' other descendents. Clearly, this probability is proportional to $\rho(v' + 1)$ and to the degree $v' + 1$, with the result

$$\rho_1(v') = \frac{v' + 1}{v_0}\ \rho(v' + 1) \quad . \tag{1.5}$$

The generating function of ρ_1 is thus given by

$$G_1(x) = \sum_{v' \geq 0} \rho_1(v')\ x^{v'} = \frac{1}{v_0}\ \frac{dG_0}{dx}(x) \quad ; \tag{1.6}$$

notice that $G_1(1)$ by the mere definition of v_0, and ρ_1 is normalised as expected[2].

1.2 Typical properties of Poissonian random graphs

1.2.1 *Overview of rigorous results*

We start with some vocabulary. A (connected) component of a graph G is called cluster. The size (order) of a cluster is the number of vertices it contains. An isolated vertex is a cluster of size unity. The number of components of G is

[2] For the K–regular and the Poissonian graphs, these generating functions are, respectively, equal to $G_0(x) = x^K, G_1(x) = x^{K-1}$, and $G_0(x) = G_1(x) = \exp(-c + cx)$. The equality between G_0 and G_1 in the latter case reflects the absence of correlation between edges in the Poissonian distribution.

denoted by $\Phi(G)$. We shall indicate its normalized fraction, $\Phi(G)/N$, by $\varphi(G)$, which is bounded from above by unity, and from below by e^{-c}, the fraction of isolated vertices. Components may be sorted by decreasing sizes $\Gamma_i(G)$, from the largest ($i = 1$) to the smallest one ($i = \Phi(G)$).

Erdös and Rényi were able to prove the following results on the sizes of the largest components:

- When $c < 1$, the largest clusters include with high probability a number of vertices scaling asymptotically as,
$$\Gamma_i(G) \sim \Theta(\ln N) \qquad\qquad (c < 1) \ , \tag{1.7}$$
for all $1 \le i \le I$, with I finite as N gets large. Most components include only a bounded number vertices. More precisely the number of components (divided by N) with a bounded number Γ of vertices is almost surely equal to
$$J(\Gamma, c) = \frac{1}{c}\frac{\Gamma^{\Gamma-2}}{\Gamma!}\left(c\,e^{-c}\right)^{\Gamma} \quad . \tag{1.8}$$
- At $c = 1$, the largest components contain $O(N^{2/3})$ vertices. More precisely, there exists two positive constants α_1, α_I such that
$$\alpha_I\, N^{2/3} < \Gamma_I(G) \le \Gamma_{I-1}(G) \le \ldots \le \Gamma_1(G) \le \alpha_1\, N^{2/3} \qquad (c = 1), \tag{1.9}$$
with I finite as N gets large.
- When $c > 1$, a drastic separation takes place between the largest cluster and all other smaller components. With high probability, $\Gamma_1(G)/N = \gamma_1(c)$ where $\gamma_1(c)$ is the unique positive solution of
$$1 - \gamma = e^{-c\,\gamma} \qquad . \tag{1.10}$$
Again, the sizes of smaller components ($i \ge 2$) are given by eqn (1.7,1.8).

The phenomenon taking place at $c = 1$ is called percolation. While below the percolation threshold $c \le 1$, most components are small with $O(\ln N)$ sizes, they percolate at the transition and give birth to a giant component with $O(N)$ vertices. Fig. 1.2A shows the fraction $\gamma_1(c)$ of vertices belonging to the giant cluster. We also report on Fig. 1.2B the average fraction of clusters,

$$\varphi^*(c) = -\frac{c}{2}\left(1 - \gamma_1(c)^2\right) + \left[1 - \gamma_1(c)\right]\left[1 - \ln\left(1 - \gamma_1(c)\right)\right] \qquad , \tag{1.11}$$

which will be useful in the following. Both are non analytic functions of their argument γ at the critical point $c = 1$.

At the phase transition, the largest components scale as $N^{2/3}$, showing that the fractions of vertices they contain scale as $N^{-1/\nu}$ with a finite size scaling exponent $\nu = 3$. The reader is referred to the existing literature for more elaborate results on the percolation transition in random graphs.

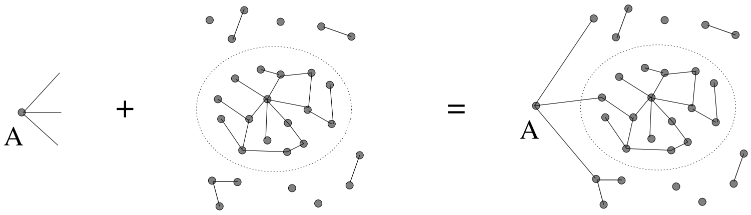

FIG. 1.3. Addition process of one vertex A and its $v = 3$ edges to a random graph G. Prior to merging, G is made of isolated nodes, small connected components, and a giant component (surrounded by the dotted line). After merging, the new vertex belongs to the giant component.

1.2.2 *Heuristic description of the giant component*

Some intuition about the above results may be gained from a simple argument. This reasoning is not mathematically correct but the results it leads to are. The starting point is an ubiquitous idea in probability and statistical physics, which could be phrased as follows: "if a system is very large, its statistical properties should be, in some sense, unaffected by a small increase in size". We now use this principle, which is merely a definition of what very large means, to determine the size γ of the giant component as a function of the connectivity c (1.10).

Consider a random graph G over N vertices, with connectivity c. Add a new vertex A to the graph to obtain G' (Fig. 1.3). If we want G' to be drawn from the same distribution as G, a number v of edges must be attached to A, where v an integer–valued random number following the Poisson distribution ρ (1.3). After addition of A, some connected components of G will merge in G'. In particular, with some probability p_v, A will not be part of the giant component of G'. To estimate p_v, we note that this event happens if and only if none of the v neighbors of A in G' belongs to the giant component of G. Thus,

$$p_v = (1-\gamma_1)^v \qquad , \tag{1.12}$$

where γ_1 is the size (fraction of vertices) of the latter. Summing both sides of (1.13) over the distribution (1.3) of v, and asserting that the change in size of the giant component between G and G' is $o(1)$ for large N, we obtain

$$1-\gamma_1 = \sum_{v\geq 0} \rho(v)\, p_v = \sum_{v\geq 0} e^{-c}\,\frac{(c(1-\gamma_1))^v}{v!} = e^{-c\,\gamma_1} \qquad , \tag{1.13}$$

which is precisely equation (1.10) for γ_1.[3]

[3] We have ruled out here the coexistence of two distinct giant components above the threshold, which would be very likely to get connected after the addition to the graph of a large (but still small with respect to N) number of vertices and attached edges.

For non-Poissonian ensembles the condition for percolation is that the first derivative of G_1 in $x = 1$ is larger than unity or, equivalently from (1.6),

$$\sum_{v \geq 0} v(v-2)\, \rho(v) > 0 \quad . \tag{1.14}$$

This equation can be simply obtained with the heuristic argument above.

1.3 Percolation transition and random 2-XORSAT

We will now illustrate the properties of random graphs on an interesting example, the random 2-XORSAT problem.

1.3.1 *Linear systems of Boolean equations*

Linear systems of Boolean equations look very much like their well known counterparts for integer-valued variables, except that equalities are defined modulo two. Consider a set of N Boolean variables x_i with indices $i = 1, \ldots, N$. Any variable shall be False (F) or True (T). The sum of two variables, denoted by +, corresponds to the logical exclusive OR between these variables defined through,

$$\begin{aligned} F + T = T + F = T \quad , \\ F + F = T + T = F \quad . \end{aligned} \tag{1.15}$$

In the following we shall use an alternative representation of the above sum rule. Variables will be equal to 0 or 1, instead of F or T respectively. Then the + operation coincides with the addition between integer numbers modulo two.

The following is a linear equation involving three variables,

$$x_1 + x_2 + x_3 = 1 \quad . \tag{1.16}$$

Four among the $2^3 = 8$ assignments of (x_1, x_2, x_3) satisfy the equation: $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$ and $(1, 1, 1)$. A Boolean system of equations is a set of Boolean equations that have to be satisfied together. For instance, the following Boolean system involving four variables

$$\begin{cases} x_1 + x_2 + x_3 = 1 \\ x_2 + x_4 = 0 \\ x_1 + x_4 = 1 \end{cases} \tag{1.17}$$

has two solutions: $(x_1, x_2, x_3, x_4) = (1, 0, 0, 0)$ and $(0, 1, 0, 1)$. A system with one or more solutions is called satisfiable. A trivial example of an unsatisfiable Boolean system is

$$\begin{cases} x_1 + x_2 + x_3 = 1 \\ x_1 + x_2 + x_3 = 0 \end{cases} \quad . \tag{1.18}$$

Determining whether a Boolean system admits an assignment of the Boolean variables satisfying all the equations constitutes the XORSAT (exclusive OR

Satisfaction) problem. In the following, we shall restrict to K-XORSAT, a variant of XORSAT where each Boolean equation include K variables precisely.

K-XORSAT belongs to the class P of polynomial problems. Determining whether a system is satisfiable or not can be achieved by the standard Gaussian elimination algorithm in a time (number of elementary operations) bounded from above by some constant times the cube of the number of bits necessary to store the system[4]

If the decision version of K-XORSAT is easy its optimization version is not. Assume you are given a system F, run the Gauss procedure and find that it is not satisfiable. Determining the maximal number $M_S(F)$ of satisfiable equations is a very hard problem. Even approximating this number is very hard. It is known that there is no approximation algorithm (unless P=NP) for XORSAT with ratio $r > \frac{1}{2}$, that is, guaranteed to satisfy at least $r \times M_S(F)$ equations for any F. But $r = \frac{1}{2}$ is achieved, on average, by making a random guess[5]!

1.3.2 *Models for random systems*

There are many different ways of generating random Boolean systems. Perhaps the simplest one is the following, called *fixed-size ensemble*. To build an equation we pick up uniformly at random K distinct indices among the N ones, say, i_1, i_2 and i_k. Then we consider the equation

$$x_{i_1} + x_{i_2} + \ldots + x_{i_k} = v \ . \tag{1.19}$$

The second member, v, is obtained by tossing a coin: $v = 0$ or $v = 1$ with equal probabilities (one half) and independently of the indices of the variables in the first member. The process is repeated M times, without correlation between equations to obtain a system with M equations.

Another statistical ensemble is the *fixed-probability ensemble*. One scans the set of all $H = 2\binom{N}{K}$ equations one after the other. Each equation is added to the system with probability p, discarded with probability $1-p$. Then a system with, on average, $p\,H$ equations (without repetition) is obtained. In practice one chooses $p = \frac{M}{H}$ to have the same (average) number of equations as in the fixed-size ensemble.

The above distributions are not the only possible ones. However they are easy to implement on a computer, are amenable to mathematical studies, and last but not least, lead to a surprisingly rich phenomenology. One of the key quantities which exhibits an interesting behaviour is

$$\begin{aligned} P_{SAT}(N,\alpha) = {} & \text{Probability that a system of random K-XORSAT with} \\ & N \text{ variables and } M = \alpha\, N \text{ equations is satisfiable ,} \end{aligned}$$

[4] The storage space is K times the number of equations times the number of bits necessary to label a variable, that is, the logarithm of the number of variables appearing in the system.

[5] Any equation is satisfied by half of the configurations of a variables, so a randomly chosen configuration satisfies on average $\frac{M}{2} \geq \frac{M_S(F)}{2}$ equations.

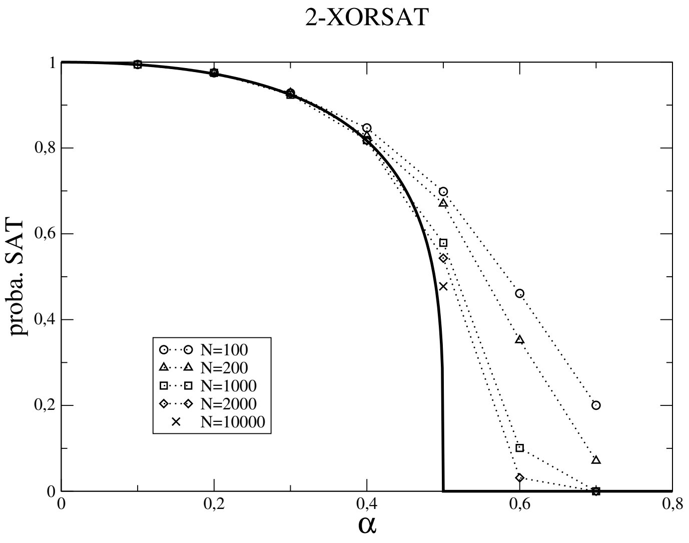

FIG. 1.4. Probability that a random 2-XORSAT formula is satisfiable as a function of the ratio α of equations per variable, and for various sizes N. The full line is the asymptotic analytical formula (1.26).

which obviously depends on K and the statistical ensemble. Given N P_{SAT} is a decreasing function of α. We will see that, in the infinite size limit (and for $K \geq 2$), the decrease is abrupt at some well defined ratio, defining a phase transition between Satisfiable and Unsatisfiable phase.

1.3.3 *Phase transition in random 2-XORSAT*

In 2-XORSAT each equation defines a joint constraint on two variables. Formulas of 2-XORSAT can be represented by a graph with N vertices (one for each variable), and αN edges. To each equation of the type $x_i + x_j = e$ corresponds an edge linking vertices i and j, and carrying 0 or 1 label (the value e of the second member). Depending on the input model chosen (Section 1.3.2) multiple edges are present or not. Clearly, the underlying graph is, for large N, is statistically Poissonian with average degree:

$$c = 2\alpha \ . \tag{1.20}$$

Figure 1.4 shows the probability P_{SAT} that a randomly extracted 2-XORSAT formula is satisfiable as function of α, and for various sizes N. It appears that P_{SAT} drops quickly to zero for large N when α reaches the percolation threshold $\alpha_c = \frac{1}{2}$. For ratios smaller than α_c the probability of satisfaction is positive, but smaller than unity.

Take $\alpha < \frac{1}{2}$. Then the random graph G associated to a random 2-XORSAT formula is non percolating, and made of many small components. Identical components (differing only by a relabelling of the variables) may appear several times, depending on their topology. For instance consider a connected graph G' made of E edges and V vertices. The average number of times G' appears in G is a function of E and V only,

$$N_{E,V} = \binom{N}{V} \left(\frac{2\alpha}{N}\right)^E \left(1 - \frac{2\alpha}{N}\right)^{\frac{V(V-1)}{2} + V(N-V)} \tag{1.21}$$

since any vertex in G' can establish edges with other vertices in G', but is not allowed to be connected to any of the $N - V$ outside vertices. When N is very large compared to E, V we have

$$N_{E,V} \simeq N^{V-E} \, \frac{(2\alpha)^E}{V!} \, e^{-2\alpha \, V} \ . \tag{1.22}$$

Three cases should distinguished, depending on the value of $V - E$:

- $V - E = 1$: this is the largest value compatible with connectedness, and corresponds to the case of trees. From (1.22) every finite tree has of the order of N copies in G.
- $V - E = 0$: this correspond to trees with one additional edge, that is, to graphs having one cycle (closed loop). The average number of unicyclic graphs is, from (1.22), finite when $N \to \infty$.
- $V - E \leq -1$: the average number of components with more than one cycle vanishes in the large N limit; those graphs are unlikely to be found and can be ignored[6].

Obviously a 2-XORSAT formula with tree structure is always satisfiable[7]. Hence dangerous subformulas, as far as satisfiability is concerned, are associated to unicyclic graphs. A simple thought shows that a unicyclic formula is satisfiable if and only if the number of edges carrying label 1 along the cycle is even. Since the values attached to the edges (second members in the formula) are uncorrelated with the topology of the subgraph (first members) each cycle is satisfiable with probability one half. We end up with the simple formula

$$P_{SAT}(N, \alpha) = \langle 2^{-C(G)} \rangle \tag{1.23}$$

where $C(G)$ denotes the number of cycles in G, and $\langle . \rangle$ the average over G. For a reason which will become clear below let us classify cycles according to their length L. How many cycles of length L can we construct? We have to choose first L vertices among N, and join them one after the order according to some order. As neither the starting vertex nor the direction along the cycle matter, the average number of L-cycles is

[6] The probability that such a graph exists is bounded from above by the average number.

[7] Start from one leaf, assign the attached variable to 0, propagate to the next variable according to the edge value, and so on, up to the completion of the tree.

$$N_L = \frac{N(N-1)\dots(N-L+1)}{2L} \times \left(\frac{2\alpha}{N}\right)^L \to \Lambda_L = \frac{(2\alpha)^L}{2L} \ . \tag{1.24}$$

when $N \to \infty$. As the emergence of a cycle between L vertices is a local event (independent of the environment) we expect the number of L-cycles to be Poisson distributed in the large N limit with parameter Λ_L. This statement can actually be proven, and extended to any finite collection of cycles of various lengths[**?**]: in the infinite size limit, the joint distribution of the numbers of cycles of lengths $1, 2, \dots, L$ is the product of Poisson laws with parameters $\Lambda_1, \Lambda_2, \dots, \Lambda_L$ calculated in (1.24). The probability of satisfaction (1.23) therefore converges to

$$\lim_{N\to\infty} P_{SAT}(N,\alpha) = \prod_{L\geq L_0} \left\{ \sum_{C\geq 0} e^{-\Lambda_L} \frac{(\Lambda_L/2)^C}{C!} \right\} = \prod_{L\geq L_0} e^{-\Lambda_L/2} \tag{1.25}$$

where L_0 is the minimal cycle length. In normal random graphs $L_0 = 3$ since triangles are the shortest cycles. However in our 2-XORSAT model any equation, or more precisely, any first member can appear twice or more, hence $L_0 = 2$. We conclude that

$$\lim_{N\to\infty} P_{SAT}(N,\alpha) = e^{\alpha/2}\,(1-2\alpha)^{\frac{1}{4}} \qquad \text{when} \qquad \alpha < \alpha_c = \frac{1}{2} \ . \tag{1.26}$$

The agreement of this result with the large size trend coming out from numerical simulations is visible in Figure 1.4. As P_{SAT} is a decreasing function of α it remains null for all ratios larger than α_c. The non analyticity of P_{SAT} at α_c locates the Sat/Unsat phase transition of 2-XORSAT.

It is an implicit assumption of statistical physics that asymptotic results of the kind of (1.26), rigorously valid in the $N \to \infty$ limit, should reflect with good accuracy the finite but large N situation. An inspection of Figure 1.4 shows this is indeed the case. For instance, for ratio $\alpha = .3$, (1.26) cannot be told from the probability of satisfaction measured for formulas with $N = 100$ variables. This statement does not hold for $\alpha = .4$, where the agreement between infinite size theory and numerics sets in when $N = 1000$ at least. It appears that such finite-size effects become bigger and bigger as α gets closer and closer to the Sat/Unsat threshold. We can guess what happens right at the threshold with the following heuristic argument.

From rigorous results on critical random graphs the largest components at the percolation threshold have size $N^{\frac{2}{3}}$. As the slope of the giant component size is finite when $c \to 1^+$ we can guess that the width of the critical region of 2-XORSAT is $\Delta c \sim N^{-1/3}$. Loosely speaking it means that a formula with N variables and $\frac{N}{2}(1+\Delta c)$ equations is 'critical' when $N\,\Delta c \sim N^{\frac{2}{3}}$.

A consequence of (1.26) with $\alpha = \frac{1}{2} - a\,N^{-1/3}$, is that the probability of satisfaction decays as

$$P_{SAT}\left(N, \alpha_c - a\,N^{-1/3}\right) \sim \mu(a)\,N^{-\frac{1}{12}} \ , \tag{1.27}$$

1

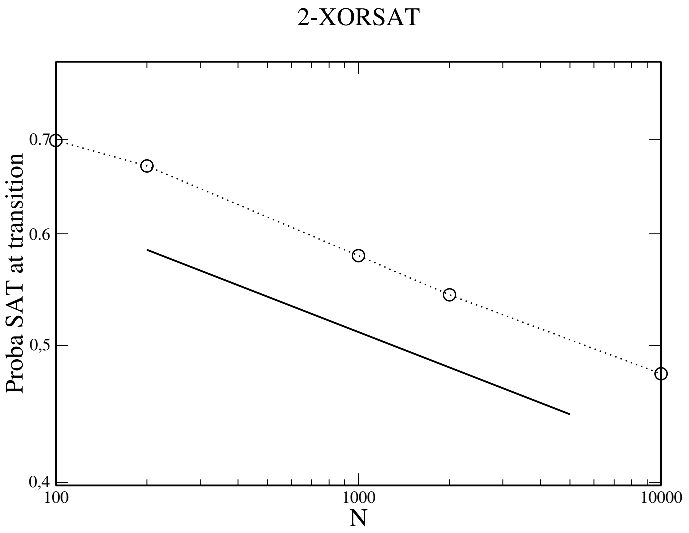

FIG. 1.5. P_{SAT} as a function of the size N at the Sat/Unsat ratio for 2-XORSAT in log-log plot. The slope $-\frac{1}{12}$ (1.27) is shown for comparison.

where μ is some function of a. This scaling agrees with numerical experiments right at the threshold ($a = 0$), though the small value of the decay exponent makes an accurate check delicate (Figure 1.5).

1.4 The Potts model and rare random graphs

1.4.1 *Relationship with Random Graphs*

In the mean-field Potts model (Potts 1952) each of the N spins $\sigma_i, i = 1, \ldots, N$ can take q values, $\sigma_i = 0, 1, ..., q-1$. The energy function reads

$$E[\sigma_1, \sigma_2, \ldots, \sigma_N] = -\frac{1}{N} \sum_{i<j} \delta_{\sigma_i, \sigma_j} \quad , \tag{1.28}$$

where $\delta_{a,b}$ is the Kronecker delta function. Note that the coupling, $\frac{1}{N}$, is chosen to ensure that the energy is of the order of N at low temperature as in the Ising case. The partition function of the Potts model is

$$Z_N(q, T) = \sum_{\{\sigma_i = 0, \ldots, q-1\}} \exp\left(\frac{1}{TN} \sum_{i<j} \delta_{\sigma_i, \sigma_j} \right) \tag{1.29}$$

where the summation runs over all q^N spin configurations. δ taking value zero or unity, the partition function may be recast in the form (Kasteleyn, Fortuin 1969, Fortuin, Kasteleyn 1972),

$$Z_N(q,T) = \sum_{\{\sigma_i\}} \prod_{i<j} \left[1 + w\,\delta_{\sigma_i,\sigma_j}\right] \qquad , \tag{1.30}$$

with

$$w = \exp\left(\frac{1}{T\,N}\right) - 1 \qquad . \tag{1.31}$$

When expanding the product appearing in (1.30), we obtain $2^{N(N-1)/2}$ terms t, each of which composed by two factors, the first one given by w raised to a power equal to the number of δs composing the second factor. We can represent graphically each term t in the expansion as a subgraph G_t of K_N. An edge connects the vertices i and j in G_t if and only if t includes the factor $\delta_{\sigma_i,\sigma_j}$. This one–to–one mapping between the terms of the expansion and and the subgraphs of K_N allows us to rewrite the partition function as

$$Z_N(q,T) = \sum_{\{\sigma_i\}} \sum_{G\subset K_N} w^{N_E(G)} \prod_{k=0}^{N_E(G)} \delta_{\sigma_{i_k},\sigma_{j_k}} \tag{1.32}$$

where $N_E(G)$ is the number of edges in the subgraph G and i_k, j_k are the vertices connected by the kth edge of the subgraph. We now exchange the order of the summations and perform the sum over the spin configurations first. Given a subgraph G with N_E edges and Φ connected components (isolated vertices included), the sum over spins configurations will give zero unless all the σs belonging to a connected component of G have the same value (as a consequence of the δ factors). In such a component, one can set the σs to any of the q different values. Hence,

$$Z_N(q,T) = \sum_{G\subset K_N} w^{N_E(G)}\, q^{\Phi(G)} \qquad . \tag{1.33}$$

As we can choose the temperature T at our convenience so can we with the weight w. Let us thus pick up a real positive number c (smaller than N), and choose $w = \frac{c}{N-c}$, that is, a temperature $T = T_p(c,N) \equiv [N \ln(N/(N-c))]^{-1}$ from (1.31). From identity (1.33) we obtain

$$\begin{aligned} Z_N\big(q, T_p(c,N)\big) &= \sum_{G\subset K_N} \left(\frac{c}{N-c}\right)^{N_E(G)} q^{\Phi(G)} \qquad (1.34) \\ &= \frac{1}{(1-\frac{c}{N})^{\binom{N}{2}}} \sum_{G\subset K_N} \left(\frac{c}{N}\right)^{N_E(G)} \left(1-\frac{c}{N}\right)^{\binom{N}{2}-N_E(G)} q^{\Phi(G)} \ . \end{aligned}$$

The key observation, due to Kasteleyn and Fortuin (1969), is that the product of the two factors depending on $N_E(G)$ in the above sum can be identified with the probability $\mathcal{P}_I(G)$ of graph G in Model I for Poissonian random graphs (1.1). In other words, up to a q-independent multiplicative factor, the partition function of the Potts model at temperature $T_p(c,N)$ is equal to the generating function of the number $\Phi(G)$ of connected components of random graphs G with average degree c,

$$Y_N(q,c) = \sum_{G \subset K_N} \mathcal{P}_I(G)\; q^{\Phi(G)} \; . \tag{1.35}$$

In particular the moments of Φ could be calculated from the knowledge of Y through successive differentiations with respect to q in the vicinity of $q=1$. A major difficulty is that, while q is a dummy variable in the generating function Y, it takes only integer values in the Potts partition function! Differentiation with respect to q requires an analytic continuation procedure we now present.

For the sake of the simplicity we hereafter restrict to the calculation of the average value of the number of components per vertex,

$$\varphi^*(c) = \lim_{N\to\infty} \sum_{G\subset K_N} \mathcal{P}_I(G)\; \frac{\Phi(G)}{N} \; , \tag{1.36}$$

to show how rigorous results presented earlier can be found back in a purely statistical physics framework. Higher moments, and the large deviations of Φ with respect to its expectation can be calculated along the same lines (Engel, Monasson 2004).

1.4.2 *Brief reminder on large deviations*

Large deviation theory is the field of probability which deals with very unlikely events. You are given a fair (unbiased) coin and toss it N times. The number H of head draws has probability

$$p_N(H) = \frac{1}{2^N} \binom{N}{H} \; . \tag{1.37}$$

When N gets large H is highly concentrated around $H^* = N/2$ with small relative fluctuations of the order of $O(\sqrt{N})$. Yet we can ask for the probability of observing a fraction $h = H/N$ equal to say, 25%, of heads, far away from the likely value $h^* = 50\%$. To calculate this probability we use Stirling's asymptotic expression for the binomial coefficient in (1.37) to obtain

$$p_N(H = h\,N) = e^{-N\omega(h)+o(N)} \; , \tag{1.38}$$

where

$$\omega(h) = \ln 2 + h \ln h + (1-h)\ln(1-h) \tag{1.39}$$

is called rate function. The meaning of (1.38) is that events with value of $h \neq h^*$ are exponentially rare in N, and $\omega(h)$ give the decay (rate) exponent. The answer to our question is $e^{-N\omega(.25)} \sim e^{-0.13\,N}$ when N is large. Some comments are:

- $\omega(h)$ is strictly positive, except in $h = h^* = \frac{1}{2}$ where it vanishes. This is the only value for the fraction of head draws with non exponentially small–in–N probability.
- Let $h = h^* + \delta h$ where δh is small. Using $\omega(h^*) = \omega'(h^*) = 0$ we have

$$P_N\big(H = (h^* + \delta h)N\big) = \exp\big[- N\, \frac{1}{2}\omega''(h^*)\, (\delta h)^2 + \dots \big] \ , \tag{1.40}$$

  that is, δh is Gaussianly distributed with zero mean and variance

$$\frac{1}{N\,\omega''(h^*)} = \frac{1}{4N} \quad . \tag{1.41}$$

  Hence central limit theorem is found back from the parabolic behaviour of the rate function around its minimum[8].
- ω is here a convex function of its argument. This property is true rate functions describing independent events. Indeed, suppose we have H positive (according to some criterion e.g. being a head for a coin) events among a set of N events, then another set of N' events among which H' are positive. If the two sets are uncorrelated

$$p_{N+N'}(H + H') \geq p_N(H) \times p_{N'}(H') \tag{1.42}$$

  since the same total number $H+H'$ of positive events could be observed in another combination of $N+N'$ events. Taking the logarithm and defining $h = H/N$, $h' = H'/N$, $u = N/(N+N')$ we obtain

$$\omega(u\, h + (1-u)\, h') \leq u\,\omega(h) + (1-u)\,\omega(h') \ , \tag{1.43}$$

  for any $u \in [0;1]$. Hence the representative curve of ω lies below the chord joining any two points on this curve, and ω is convex. Non-convex rate functions are found in presence of strong correlations[9].

1.4.3 *Large deviations for the number of components of random graphs*

To describe the decomposition of a large random graph into its components, it is convenient to introduce the probability $P(\Phi; c, N)$ of a random graph with N vertices to have Φ components

$$P(\Phi; c, N) = \sum_G \mathcal{P}_I(G)\ \delta(\Phi, \Phi(G)), \tag{1.44}$$

[8] Non standard behaviour e.g. fluctuations of the order of N^ν with $\nu \neq \frac{1}{2}$ as found in Levy flights correspond to non-analyticies of ω in h^* or the vanishing of the second derivative.

[9] Consider the following experiment. You are given three coins: the first one is fair (coin A), the second and third coins, respectively denoted by B and C, are biased and give head with probabilities, respectively, $\frac{1}{4}$ and $\frac{3}{4}$. First draw coin A once. If the outcome is head pick up coin B, otherwise pick up coin C. Then draw your coin N times. What is the rate function associated to the fraction h of heads?

where $\delta(a,b)$ denotes the Kronecker delta.

A general observation is that for given c and large N the probability $P(\Phi;c,N)$ gets sharply peaked at some *typical* value Φ^* of Φ, and the probabilities for values of Φ significantly different from Φ^* being exponentially small in N. To describe this fact more quantitatively we introduce the number of components per vertex $\varphi = \Phi/N$ together with the quantity

$$\omega(\varphi,c) = \lim_{N\to\infty} \frac{1}{N} \log P(\Phi;c,N). \tag{1.45}$$

Clearly $\omega(\varphi,c) \leq 0$ and the typical value φ^* of φ has $\omega(\varphi^*,c) = 0$. Averages with $\mathcal{P}_I(G)$ are therefore dominated by graphs G with a typical number of components.

The focus of the present paper is on properties of random graphs which are *atypical* with respect to their number of components C. In order to get access to the properties of these graphs we introduce the *biased* probability distributions

$$\mathcal{P}_I(G;q) = \frac{1}{Y_N(q,c)}\, \mathcal{P}_I(G)\, q^{\Phi(G)}, \tag{1.46}$$

with $Y_N(q,c)$ defined by (1.35). Contrary to averages with $\mathcal{P}_I(G)$ those with $\mathcal{P}_I(G;q)$ are dominated by graphs with an atypical number of components which is fixed implicitly with the parameter q. Values of q smaller than 1 shift weight to graphs with few components whereas for $q > 1$ graphs with many components dominate the distribution. The typical case is obviously recovered for $q = 1$.

Similar to $\omega(\varphi,c)$ it is convenient to introduce the function

$$y(c,q) = \lim_{N\to\infty} \frac{1}{N} \log Y_N(q,c). \tag{1.47}$$

From (1.35) and (1.45) it follows to leading order in N that

$$Y_N(q,c) \simeq N \int_0^1 d\varphi\ \exp(N[\omega(\varphi,c) + \varphi \log q]) \tag{1.48}$$

and performing the integral by the Laplace method for large N we find that $y(c,q)$ and $\omega(\varphi,c)$ are Legendre transforms of each other:

$$\begin{aligned} y(c,q) &= \max_{\varphi}\, [\omega(\varphi,c) + \varphi \log q] \\ \omega(\varphi,c) &= \min_{q}\, [y(c,q) - \varphi \log q] \\ q &= \exp(-\frac{\partial \omega}{\partial \varphi})\ , \quad \varphi = q\frac{\partial y}{\partial q} \end{aligned} \tag{1.49}$$

The large deviation properties of the ensemble of random graphs as characterized by $\omega(\varphi,c)$ can hence be inferred from $y(c,q)$. For more information, see A. Engel,

R. Monasson, A.K. Hartmann, On large-deviation properties of Erdos-Renyi random graphs, Journal of Statistical Physics 117, 387 (2004).

In the next section we show how $y(c,q)$ can be obtained from the statistical mechanics of the Potts model.

1.5 Statistical mechanics of the Potts model

1.5.1 *Free-energy in the non-percolating phase.*

Let us call $f(q,T)$ the density of free energy of the mean-field Potts model in the large N limit. We specialise in the following to the temperature $T_p(c, N \to \infty) = \frac{1}{c}$. Our aim is to obtain an analytic continuation of f to real-valued q, allowing us to calculate the average number of components[10]

$$\varphi^*(c) = -c \, \frac{\partial f}{\partial q} \left(1, \frac{1}{c} \right) \quad . \tag{1.51}$$

The energy function (1.28) depends on the configuration of spins $\mathcal{C} = \{\sigma_i\}$ through the fractions $x(\sigma;\mathcal{C})$ of variables σ_i equal to σ $(= 0, 1, \cdots, q-1)$ (Wu 1982),

$$x(\sigma;\mathcal{C}) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\sigma_i,\sigma}, \quad (\sigma = 0, 1, ..., q-1) \qquad . \tag{1.52}$$

Of course the sum of the fractions (1.52) over σ (for a given $\mathcal{C}$) is equal to unity. Energy (1.28) may be rewritten in terms of the xs as

$$E[\mathcal{C}] = -\frac{N}{2} \sum_{\sigma=0}^{q-1} \left[x(\sigma;\mathcal{C}) \right]^2 + \frac{1}{2} \qquad . \tag{1.53}$$

The last term on the r.h.s. of (1.53) can be neglected with respect to the first term whose order of magnitude is $O(N)$. The sum over the q^N spin configurations in the partition function (1.29) may be replaced with a sum over the value of the fractions $x(\sigma)$,

$$Z_N\left(q, \frac{1}{c}\right) = \sum_{\{x(\sigma)=0,\frac{1}{N},\frac{2}{N},\ldots,1\}}^{(R)} D_N[\{x(\sigma)\}] \, \exp\left(\frac{cN}{2} \sum_{\sigma=0}^{q-1} x(\sigma)^2 \right) \tag{1.54}$$

[10] From definitions (1.35,1.36) $\varphi^*(c)$ is the partial derivative of the generating function $Y_N(q,c)$ in $q = 1$, divided by N. As $Y_N(1,c) = 1$ we can differentiate $\ln Y_N(q,c)$ instead of $Y_N(q,c)$ itself. Kasteleyn-Fortuin correspondence says that $\ln Z_N(q, \frac{1}{c}) = \ln Y_N(q,c)$ up to an additive q-independent constant. Thus

$$\varphi^*(c) = \lim_{N\to\infty} \frac{1}{N} \, \frac{\partial \ln Z_N}{\partial q} \left(1, \frac{1}{c}\right) . \tag{1.50}$$

Permutation of the limit and the differentiation operations, and insertion of the definition of the free energy density lead to (1.51).

where the multiplicity D_N of (number of spin configurations associated to) a set of fractions $\{x(\sigma\}$ is

$$D_N[\{x(\sigma\}] = \binom{N}{N\,x(0),\ldots,N\,x(q-1)} \ . \tag{1.55}$$

The subscript (R) indicates that the sum (1.54) is restricted to the normalized subspace of the q-dimensional positive hypercube,

$$(R) \qquad \sum_{\sigma=0}^{q-1} x(\sigma) = 1 \qquad . \tag{1.56}$$

In the limit of large N the sum in (1.35) coincides, to exponential order in N, with its largest term. The Potts free energy density reads

$$f\left(q,\frac{1}{c}\right) = \overset{(R)}{\min_{\{x(\sigma)\}}}\ \hat{f}[\{x(\sigma)\}] \tag{1.57}$$

where $\hat{f}$ is the q-multivariate function,

$$\hat{f}\big[x(0),x(1),\ldots,x(q-1)\big] = \sum_{\sigma=0}^{q-1}\left\{ -\frac{1}{2}x(\sigma)^2 + \frac{1}{c}\,x(\sigma)\ln x(\sigma)\right\} \quad , \tag{1.58}$$

and the minimum is sought under constraint (1.56).

Given the initial formulation of the problem each value of σ among $0,\ldots,q-1$ plays the same role; indeed $\hat{f}$ is invariant under any permutation of its q arguments. Consequently, if $\{x(\sigma)\}$ is a minimum of $\hat{f}$, so is $\{x(\pi\sigma)\}$ for any permutation π of the symmetric group S_q. We shall see that, depending on the value c of the average degree, the permutation symmetry may, or may not be broken. The breaking of the permutation symmetry of the q fractions (in the limit $q \to 1$!) coincides with the birth a giant component in the associated random graph.

Consider first the permutation symmetric (PS) hypothesis for which the minimum is unique. Then all fractions have a common value equal to, from (1.56),

$$x^{PS}(\sigma) = \frac{1}{q}, \qquad \forall\, \sigma = 0,\ldots,q-1. \tag{1.59}$$

Inserting (1.59) into the Potts free energy function (1.58) we obtain the following expression for the free energy density

$$f^{PS}\left(q,\frac{1}{c}\right) = -\frac{1}{2q} - \frac{1}{c}\,\ln q \quad . \tag{1.60}$$

Though the above expression has been derived for positive integer q it can be now straightforwardly continued to real valued q, and we obtain the average number of components from (1.51),

$$\varphi^{*,PS}(c) = 1 - \frac{c}{2} \ . \tag{1.61}$$

Comparison with the rigorous result (1.11) from random graph theory indicates that the symmetric expression (1.61) is exact as long as $c \leq 1$, and is false above the percolation threshold. The breakdown of PS in the presence of a giant component is a proof for the necessity of permutation symmetry breaking.

The same conclusion can be reached from an analysis of the curvature of the free energy function (1.58) around the symmetric solution (1.59). The $q \times q$ Hessian matrix

$$\mathcal{M}^{PS}_{\sigma,\tau} = \frac{\partial^2 \hat{f}}{\partial x(\sigma)\partial x(\tau)}\big[\{x^{PS}\}\big] = \Big(\frac{q}{c} - 1\Big)\,\delta_{\sigma,\tau} \quad , \tag{1.62}$$

is diagonal . The normalised subspace (1.56) is spanned by $q-1$ eigenvectors, with the $q-1$-fold degenerate eigenvalue $\Lambda^{PS} = \frac{q}{c} - 1$. In the $q \to 1$ limit Λ^{PS} is positive for $c < 1$, and the PS Ansatz (1.59) is a true minimum of $\hat{f}$. For $c > 1$ *i.e.* above the percolation threshold the PS Ansatz is a local maximum of $\hat{f}$. This scenario is similar to the continuous phase transition taking place in the mean-field Ising model.

1.5.2 *Permutation Symmetry Breaking, Giant and Small Components*

The simplest permutation symmetry broken (PSB) Ansatz is a set of fractions in which one among the q spin values, say, $\sigma = 0$, is more frequent than the other values,

$$\begin{aligned} x^{PSB}(0) &= \frac{1}{q}\big[1 + (q-1)\,s\big] \ , \\ x^{PSB}(\sigma) &= \frac{1}{q}[1 - s] \ , \qquad (\sigma = 1, ..., q-1), \end{aligned} \tag{1.63}$$

which fulfills constraint (1.56). Parameter s may *a priori* take any real value, which enlarges the space in which a minimum can be searched with respect to the symmetric case $s = 0$ (1.59). The free energy density of the Potts model is obtained by inserting fractions (1.63) into (1.58). To obtain the average number of components $\varphi^*(c)$, it is sufficient to expand f around $q = 1$,

$$f^{PSB}\left(q, \frac{1}{c}\right) = -\frac{1}{2} + (q-1)\ \min_{s \geq 0} f_1^{PSB}(s,c) + O\big((q-1)^2\big) \tag{1.64}$$

with

$$f_1^{PSB}(s,c) = \frac{1}{2}(1 - s^2) - \frac{1}{c}\,(1-s)\big(1 - \ln(1-s)\big) \quad . \tag{1.65}$$

Minimization of $f_1^{PSB}(s,c)$ with respect to s shows that, for $c \leq 1$, the symmetric solution $s=0$ is the only one, whereas, for $c>1$, there exists a non vanishing optimal value $s^*(c)$ of s that is solution of the implicit equation

$$1 - s^* = e^{-c\, s^*} \qquad . \tag{1.66}$$

Comparing eqns (1.66) and (1.10) allows to unveil the meaning of $s^*(c)$: it is simply the fraction of vertices $\gamma_1(c)$ belonging to the giant component. Note that the average fraction of connected components, $\varphi^{*,PSB}(c) = -c\, f_1^{PSB}(s^*(c),c)$ from (1.51,1.64), agrees with (1.11)[11].

To understand the structure of the PSB minimum (1.63), and why s^* coincides with the size of the giant component, we need a multivariate generating function accounting for the size of the components, and not only their number. This can be done through the introduction of a field h into the Potts Hamiltonian (1.28),

$$E[\sigma_1,\sigma_2,\ldots,\sigma_N] = -\frac{1}{N}\sum_{i<j}\delta_{\sigma_i,\sigma_j} - h\sum_i \delta_{\sigma_i,0}. \tag{1.67}$$

This field has a tendency to push spins in the 0 value if $h>0$, or in any of the remaining $q-1$ values if $h<0$. All the calculations exposed above can be repeated, with the free energy functional (1.58) added a $-h\,x(0)$ contribution. In the infinite size limit the average fractions x may be derived from the density of free energy $f(q,T,h)$,

$$x(0) = \frac{\partial f}{\partial h}(q,T,h)\ , \qquad\qquad x(\sigma) = \frac{1-x(0)}{q-1} \qquad (\sigma \geq 1) \qquad . \tag{1.68}$$

Notice that, in presence of a non zero field, the permutation symmetry of the free energy functional $\hat{f}$ is explicitly broken. It is therefore justified, and even necessary, to look for a minimum of the form (1.63).

The one–to–one correspondence between the expansion of partition function associated to the energy function (1.67) and the subgraphs of K_N exposed in Section 1.4.1 now demands some multiplicative factor $\exp(h\,\delta_{\sigma_i,0}/T)$ to be associated to each vertex i of the subgraphs. As a consequence, the partition function reads

[11] In addition the analysis of the Hessian matrix of $\hat{f}$ around (1.63) shows the PSB Ansatz is a local minimum. Define $\mathcal{M}^{PSB}$, the $(q-1)\times(q-1)$ Hessian matrix of

$$\hat{f}\left(1-\sum_{\sigma=1}^{q-1} x(\sigma), x(1), \ldots, x(q-1)\right) .$$

Then $\mathcal{M}^{PSB}_{\sigma,\tau} = A + B\,\delta_{\sigma,\tau}$ with $A = -1+1/c/x(0)$, $B = -1+1/c/x(1)$. The eigenvalues of the Hessian matrix are: $\Lambda_1^{PSB} = B - A$, $\Lambda_2^{PSB} = B$ with degeneracy $q-2$. In the $q \to 1$ limit we find $0 < \Lambda_2 = -1 + 1/c/(1-s^*(c)) < \Lambda_1 = s^*(c)/c/(1-s^*(c))$ for any $c > 1$.

$$Z_N(q,T,h) = \sum_{G \subset K_N} w^{N_E(G)} \prod_{j=1}^{\Phi(G)} \left[q - 1 + \left(e^{h/T}\right)^{\Gamma(G_j)} \right] \qquad , \tag{1.69}$$

where the G_js denote the components of G. To calculate $x(0)$ from (1.68) we differentiate partition function (1.69) at a finite field, multiply by $-T/N/Z_N(q,T,h)$, send $N \to \infty$, and later make $h \to 0^+$. The output of this procedure is[12]

$$x(0) = \frac{1-\gamma_1}{q} + \gamma_1 \qquad , \tag{1.70}$$

where $\gamma_1 = \Gamma_1/N$ is the average fraction of vertices in the giant component (if any). Comparing eqns (1.70) and (1.63), we see that s parametrising the broken minimum indeed coincides with γ_1.

The whole distribution of component sizes is easy to derive from this approach (Lubensky, McKane, 1981). First we differentiate eqn (1.69) with respect to h, then send $N \to \infty$, and finally send $q \to 1$ to obtain

$$\sum_{C=1}^{\infty} J(\Gamma,c)\,\Gamma\, e^{-h\Gamma/T} = 1 - s^*(h,c) \quad , \tag{1.71}$$

where $J(\Gamma,c)$ is the typical number per vertex of components of size C, and $s^*(h,c)$ is the location of the maximum of $-c\, f_1^{PSB}(s,c) - h(1-s)$ where the expression of f_1^{PSB} is given in (1.65). $s^*(h,c)$ is the largest root of

$$1 - s = e^{-c\,s-h} \quad . \tag{1.72}$$

Using Lagrange inversion theorem[13] we obtain s as a function of h, and identify the power of $e^{-h/T}$ on both sides of eqn (1.71) to obtain (1.8) for any $\Gamma \geq 1$.

1.5.3 *At the critical point*

Consider the case of critical random graphs, where the average degree is $c = 1$. Then the free energy f_1^{PSB} is given by

$$f_1^{PSB}(s,1) = \frac{1}{2}(1-s^2) - (1-s)\big(1 - \ln(1-s)\big) = -\frac{1}{2} + \frac{s^3}{6} + O(s^4) \quad . \tag{1.73}$$

Hence, for finite sizes N, the distribution of the order parameter s, which corresponds to the number of vertices in the giant component, is given by

[12] It is assumed here that there is one component G_1 of size $O(N)$, and $N - o(N)$ components of size $O(1)$. This is only assumption stable against edge-addition, see heuristic argument to calculate the size of the giant component.

[13] Let us define $t = (1-s)/c$, $u = c\exp(-c-h)$, then $u = t\varphi(t)$ where $\varphi(t) = \exp(-t)$. Inversion of $u(t)$ gives $t(u) = \sum_{n\geq 1} t_n u^n/n!$ where t_n is the $n-1$th derivative of $\varphi(t)^n$ in $t = 0$.

$$P(s) \propto \exp\left(- \frac{N}{6}(q-1)s^3 \right) \quad . \tag{1.74}$$

Hence we expect $s \sim N^{-1/3}$, that is, large components with $O(N^{2/3})$ vertices. A more precise calculation can be done with a non-zero field h to obtain the accurate distribution of the largest component size. This distribution has been calculated exactly by B. Pittel, On the largest component of a random graph at a nearcritical stage, Journal of combinatorial theory, B82, 237 (2000).

1.6 Exercise

Consider the random 1-XORSAT model with N variables, and ratio α of clauses per variables.

1. Show that the critical value of the ratio above which the probability of satisfaction vanishes when $N \to \infty$ is $\alpha_c = 0$.

2. Show that, for $\alpha > 0$, there exists a strictly positive rate function defined through

$$\lim_{N\to\infty} \frac{1}{N} \log P_{SAT}(N, \alpha) = -\omega(\alpha) \ . \tag{1.75}$$

Calculate $\omega(\alpha)$ in the fixed-size ensemble, and then in the fixed-probability ensemble.