Statistical mechanics perspective on the phase transition in vertex covering of finite-connectivity random graphs

Alexander K. Hartmann and Martin Weigt
Institute for Theoretical Physics,
University of Göttingen, Bunsenstr. 9, 37073 Göttingen, Germany
E-mail: hartmann/weigt@theorie.physik.uni-goettingen.de

February 1, 2008

Abstract

The vertex-cover problem is studied for random graphs $G_{N,cN}$ having N vertices and cN edges. Exact numerical results are obtained by a branch-and-bound algorithm. It is found that a transition in the coverability at a c-dependent threshold $x = x_c(c)$ appears, where xN is the cardinality of the vertex cover. This transition coincides with a sharp peak of the typical numerical effort, which is needed to decide whether there exists a cover with xN vertices or not. Additionally, the transition is visible in a jump of the backbone size as a function of x.

For small edge concentrations $c \ll 0.5$, a cluster expansion is performed, giving very accurate results in this regime. These results are extended using methods developed in statistical physics. The so called annealed approximation reproduces a rigorous bound on $x_c(c)$ which was known previously. The main part of the paper contains an application of the replica method. Within the replica symmetric ansatz the threshold $x_c(c)$ and the critical backbone size $b_c(c)$ can be calculated. For $c < e/2$ the results show an excellent agreement with the numerical findings. At average vertex degree $2c = e$, an instability of the simple replica symmetric solution occurs.

1 Introduction

According to Garey and Johnson [6], the vertex cover (VC) problem belongs to the six basic NP-complete problems. Here VC is investigated for an ensemble of random graphs $G_{N,cN}$ having N vertices and cN edges [4], with c constant. Despite some efforts in the past [7, 5], no solution for the critical cardinality $X_c(c)$ of the vertex cover as a function of c has been found, but some lower and upper bounds were obtained. In this paper we investigate the problem with an exact branch-and-bound algorithm, a cluster expansion for small c and with methods borrowed from the statistical physics of disordered systems [15], see also [23].

Our main result is the following, with W being the LambertW-function ($x = W(x)e^{W(x)}$)): *In the large-N limit and for $c \leq e/2$ (e Eulerian constant), the cardinality $X_c(c)$ of the minimal vertex cover of a random graph $G_{N,cN}$ is given by*

$$X_c(c) = N - \frac{2W(2c) + W(2c)^2}{4c} N + o(N) \ , \tag{1}$$

and the number of vertices being in the backbone (see below) of these minimal VCs reads

$$B_c(c) = N - \frac{W(2c)^2}{2c} N + o(N) \ . \tag{2}$$

For $c > e/2$, the expression given on the right-hand side of (1) provides a lower bound on $x_c(c)$.

The *backbone* is defined as follows: Usually for a graph different minimal vertex covers exist. A vertex which belongs either to all vertex covers or to no vertex cover of a given graph is said to belong to the backbone.

Statistical mechanics methods were already applied to other famous NP-complete problems, as *e.g.* K-satisfiability (KSAT) [17] or number partitioning [14]. They are known to show interesting phase transitions in their solvability and, even more interestingly, in their typical case algorithmic complexity, *i.e.* in the dependence of the median solution time on the system size [20, 10]. Consider *e.g.* the satisfiability problem with the number of constraints per variable as a parameter. When this parameter exceeds a certain threshold, the solvability of a randomly chosen logical formula undergoes a sharp transition from almost always satisfiable to almost always unsatisfiable [16]. The hardest to solve formulae are found in the vicinity of the transition point. Far away from this point the solution time is much smaller, as the problem is easily fulfilled or hopelessly over-constrained. The typical solution times in the under-constrained phase are even found to depend only polynomially on the system size! Recently, insight coming from a statistical-physics perspective on these problems [17] has lead to a fruitful cooperation with computer scientists, and has shed some light on the nature of this transition [19]. Frequently, on the cost of not being mathematically rigorous, methods of statistical physics allow to obtain more insight than classical tools of computer science or discrete mathematics. This is true for the VC problem as well, as will be shown in this work.

The paper is organized as follows. After this introductory section, the investigated model, related problems, and several notations are introduced. Some previously known rigorous bounds for the minimum cardinality of the vertex-cover are cited. In the third chapter VC is studied numerically with an exact branch-and-bound procedure. Then a cluster expansion for disconnected graphs with low average vertex degree is performed. Section 5 contains the main part of the paper: statistical physics strategies are applied. A short introduction is given, which relates several elements of graph theory to corresponding quantities appearing in physics. Then, two approaches are presented. The *annealed approximation* reproduces one of the above-mentioned rigorous bounds. More detailed insight is gained by the *replica method*. Using the replica symmetric ansatz, the threshold and the backbone size at the threshold can be calculated. The results are compared with the data obtained by the branch-and-bound method. In the last section conclusions and an outlook are given.

2 The model

2.1 Vertex cover and related problems

In this section we want to introduce the investigated model.

Take any graph $G = (V, E)$ with the N vertices $i \in \{1, ..., N\}$ and M edges $(i, j) \in E \subset V \times V$. A *vertex cover* (VC) is a subset $V_{VC} \subset V$ of vertices such that for every edge $(i, j) \in E$ there is at least one of its endpoints i or j in V_{VC}. We call the vertices in V_{VC} covered, whereas the vertices in its complement $V \setminus V_{VC}$ are called uncovered.

Also *partial covers* are considered. In this case the set V_{VC} is not a VC and there are some edges (i, j) with $i \notin V_{VC}$ and $j \notin V_{VC}$. In this case we call the edge uncovered as well. The task of finding the minimum number of uncovered edges given a graph G and the cardinality $X \equiv |V_{VC}|$ is an optimization problem.

The corresponding *decision problem*, whether there exists a VC V_{VC} of fixed cardinality $X = |V_{VC}|$, with $1 \leq X < N$, is according to Garey and Johnson [6] one of the six basic

NP-complete problems. So it is widely believed that one cannot construct any algorithm which solves the problem substantially faster than exhaustive search, *i.e.* only algorithms are known which have an exponential worst-case time complexity in N and M.

VC is related to other well-known and widely used NP-complete problems. The first one is the *independent set* (ISET) problem. An ISET is a subset $V_{ISET} \subset V$ of vertices such that for all $i, j \in V_{ISET}$ we have $(i,j) \notin E$. So $V \setminus V_{ISET}$ is obviously a VC for every ISET V_{ISET}, and every maximal ISET is the complement of a minimal VC. The *independence number*, defined as the maximum of cardinalities $|V_{ISET}|$ of all ISETs, is consequently given by $N - \min_{\text{VC}} |V_{VC}|$.

A *clique* is a fully connected subgraph. So, if the subset $V_{ISET} \subset V$ is an ISET in $G = (V,E)$, it is a clique in the complementary graph $\overline{G} = (V, V \times V \setminus E)$. Finding the largest clique in one graph is equivalent to finding the largest ISET in the complementary graph.

2.2 Random graphs

In order to speak of median or average cases, and of phase transitions, we have to introduce a probability distribution over graphs. This can be done best by using the concept of *random graphs* as already introduced about 40 years ago by Erdös and Rényi [4]. A random graph $G_{N,M}$ is a graph with N vertices $V = \{1, ..., N\}$ and M randomly drawn edges such that any two instances (for fixed N, M) are equiprobable.

An alternative description would be, to include an arbitrary pair of vertices with a certain probability p. For large N, the number of edges becomes almost surely $pN^2/2 + O(N)$, and both concepts can be identified by choosing $p = 2M/N^2$.

The regime we are interested in are *finite connectivity graphs* where the average vertex degree $2c = 2M/N$ stays constant in the large N limit. Under this scaling of the edge number, the cardinality of the minimal VC should typically depend linearly on N as well, $\min_{\text{VC}} |V_{VC}| = x_c(c)N$. The main purpose of this paper is to show evidence that there is an asymptotically ($N \to \infty$) sharp threshold $x_c(c)$ which depends for almost all graphs only on the average vertex degree $2c$, and to find its functional dependence on c.

Here we want to review shortly some of the fundamental results on random graphs which were already described in [4], and which are important for the following sections:

The first point we want to mention is the distribution of vertex degrees d, in the limit $N \to \infty$ it is given by a Poisson-distribution with mean $2c$:

$$Po_{2c}(d) = e^{-2c} \frac{(2c)^d}{d!} \ . \tag{3}$$

A second point which is important for the understanding of the following is the component structure. For $c < 1/2$, *i.e.* if the vertices have in average less than one neighbor, the graph $G_{N,cN}$ is built up from connected components which have up to $O(\log N)$ vertices. The probability that a component is a specific tree T_k of k vertices is given by

$$\rho(k) = e^{-2ck} \frac{(2c)^{k-1}}{k!} \ , \tag{4}$$

and is equal for all k^{k-2} distinct trees. As the fraction of vertices which are collected in finite trees is $\sum_{k=1}^{\infty} \rho(k) k^{k-2} k = 1$ for all $c < 1/2$, in this case almost all vertices are collected in such trees. For $c > 1/2$ a giant component appears which contains a finite fraction of all vertices. $c = 1/2$ is therefore called the *percolation threshold*.

2.3 Rigorously known bounds

In this subsection we are going to present some previously known rigorous bounds on $x_c(c)$. A general one for arbitrary, *i.e.* non-random graphs was given by Harant [9] who generalized an old result of Caro and Wei [3]. Translated into our notation, he showed that

$$x_c(G) \leq 1 - \frac{1}{N} \frac{\left(\sum_{i\in V} \frac{1}{d_i+1}\right)^2}{\sum_{i\in V} \frac{1}{d_i+1} - \sum_{(i,j)\in E} \frac{(d_i-d_j)^2}{(d_i+1)(d_j+1)}} \tag{5}$$

where d_i is the vertex degree of vertex i. Using the distribution (3) of vertex degrees and its generalization to pairs of connected vertices, this can easily be converted into an upper bound on $x_c(c)$ which holds almost surely for $N \to \infty$.

The vertex cover problem or the above-mentioned related problems were also studied in the case of random graphs, and even completely solved in the case of infinite connectivity graphs, where any edge is drawn with finite probability p, such that the expected number of edges is $p\binom{N}{2} = 0(N^2)$. There the minimal VC has cardinality $(N - 2\log_{1/(1-p)} N - O(\log\log N))$ [2]. Bounds in the finite-connectivity region of random graphs with N vertices and cN edges were given by Gazmuri [7]. He showed that

$$x_l(c) < x_c(c) < 1 - \frac{\log 2c}{2c} \tag{6}$$

where the lower bound is given by the unique solution of

$$0 = x_l(c)\log x_l(c) + (1-x_l(c))\log(1-x_l(c)) - c(1-x_l(c))^2\ . \tag{7}$$

As we will see later on, this bound coincides with the so-called annealed bound in statistical physics. The correct asymptotics for large c was given by Frieze [5]:

$$x_c(c) = 1 - \frac{1}{c}(\log c - \log\log 2c + 1) + o(\frac{1}{c})\ . \tag{8}$$

3 Numerical evidence for a phase transition

To achieve a thorough insight into the nature of the problem, numerical simulations were performed. At first the branch-and-bound algorithm is explained which was implemented for this purpose. Then, results are presented which relate the transition in solvability to a change in the median-case time complexity. Also the dependence of the backbone (see below) on the cover size x shows a jump at this transition.

3.1 The algorithm

All numerical results were obtained by an exact enumeration. Using a branch-and-bound algorithm similar to [12, 21, 22] all covers can be calculated: as each vertex is either covered or uncovered, there are 2^N possible configurations which can be arranged as leafs of a binary (backtracking) tree. At each node, the two subtrees represent the subproblems where the corresponding vertex is either covered or uncovered. The *branch* operation tries to find a solution by investigating both subtrees and keeping only the optimum solutions.

First we concentrate on the algorithm which finds the configurations with the minimum number of uncovered edges for a given graph and a given number X of vertices which can be covered. We want to omit subtrees which for sure contain no optimum solutions: this is the case either if the number of covered vertices exceeds X or if the leafs of the subtree can already be proven to be worse than previously considered configurations. Thus, it is possible to avoid branching into some subtrees by calculating the following *bound*: it uses the *current* vertex degree $d(i)$, which is the number of uncovered neighbors at a specific stage of the calculation. By covering a vertex i the total number of uncovered edges is reduced by exactly $d(i)$. If several vertices $j_1, j_2, \ldots, j_k$ are covered, the number of uncovered edges is *at most* reduced by $d(j_1)+d(j_2)+\ldots+d(j_k)$. Assume that at a certain stage within the backtracking tree, there are *uncov* edges uncovered and still k vertices to cover. Then a lower bound M for the best solution which can be found in the subtree is

$$M = \max\left[0, \mathit{uncov} - \max_{j_1,\ldots,j_k} d(j_1)+\ldots+d(j_k)\right] \ . \tag{9}$$

The maximum is easily calculated by always storing the uncovered vertices sorted according their current degrees. The algorithm can avoid branching into a subtree if M is strictly larger than the number *opt* of uncovered edges in the best solution found so far. If one is interested only in an arbitrary minimum configuration instead of enumerating all, one can omit every subtree with $M \geq \mathit{opt}$. In the latter case the algorithm can be stopped as soon as a configuration with $\mathit{opt} = 0$ is found.

For the order the vertices are selected to be (un-)covered within the algorithm, the following heuristic is applied: the order of the vertices is given by their current degree. Thus, the first descent into the tree is equivalent to the greedy heuristic which iteratively covers vertices by always taking the vertex with the highest current degree. Later, it will be become clear from the results that this heuristic is indeed a suitable strategy.

The following representation summarizes the algorithm for enumerating all configurations exhibiting a minimum number of uncovered edges. Let $G = (V, E)$ be a graph, k the number of vertices to cover and *uncov* the number of edges to cover. Initially $k = X$ and $\mathit{uncov} = |E|$. The variable *opt* is initialized with $\mathit{opt} = |E|$ and contains the minimum number of uncovered edges found so far. The value of *opt* is passed via call by reference. At the beginning all vertices $i \in V$ are marked as *free*. The marks are considered to be passed via call by reference as well (not shown explicitly). Additionally it is assumed that somewhere a set of (optimum) solutions can be stored.

```
algorithm min-cover($G, k, uncov, opt$)
begin
    if k=0 then {leaf of tree reached?}
    begin
        if $uncov < opt$ then {new minimum found?}
        begin
            $opt := uncov$;
            clear set of stored configurations;
        end;
        store configuration;
    end;
    if bound condition is true (see text) then
        return;
    let $i \in V$ a vertex marked as free of maximal current degree;
    mark $i$ as covered;
    $k := k - 1$;
    adjust degrees of all neighbors $j$ of $i$: $d(j) := d(j) - 1$;
    min-cover($G, k, uncov - d(i), opt$) {branch into 'left' subtree};
    mark $i$ as uncovered;
    $k := k + 1$;
    (re)adjust degrees of all neighbors $j$ of $i$: $d(j) := d(j) + 1$;
    **min-cover**$(G, k, uncov, opt)$ {branch into 'right' subtree};
    mark $i$ as *free*;
end
```

In the actual implementation, the algorithm does not descend further into the tree as well, when no uncovered edges are left. In this case the vertex covers of the corresponding subtree consist of the vertices covered so far and all possible selections of k vertices among all uncovered vertices.

Now we discuss the case of finding a true VC of minimum cardinality, where the performance of the method can be enhanced by some extensions. The algorithm is called with $k = |V|, opt = 0$ and k is passed via call by reference like *opt*. Now assume that during the execution of the algorithm a total cover ($uncov = 0$) is found and $k > 0$. Thus it is possible to cover all edges with less than the allowed number of vertices. Consequently, it is not necessary to cover additional vertices, and the value of k is set to zero. Additionally the set of configurations which was stored before is cleared. Furthermore, whenever a vertex i is marked as *uncovered*, all its neighbors j can be covered immediately, because no uncovered edge should remain. Please note that in this case the degrees of all neighbors of the neighbors j of i have to be readjusted as well. After the initial call of this modified algorithm has finished, the variable k contains the cardinality of the minimum vertex cover.

The algorithm was implemented via the help of the LEDA library [13] which offers many useful data types and algorithms for linear algebra and graph problems. Since the VC problem is NP-hard, the method exhibits an exponential worst-case time complexity. Although our algorithm is very simple, in the regime $0.5 < c < 5$ random graphs up to size $N = 100$ could be treated for all values $X \in [0, N]$. For the calculation of covers of minimum cardinality, also graphs with $N = 140$ could be considered. Please note that for $c < 0.5$ the graphs can be divided into many connected components of sizes up to $O(\log N)$. Then, in the case one is interested only in the cover of minimum cardinality, the algorithm can be applied to each component separately, yielding only a polynomial time-complexity.

3.2 Numerical results

A first evidence for a peak of the typical case complexity near the threshold was given in [8] where the problem was matched to SAT and solved with the Davis Putnam procedure. The running time was measured for graphs of size $N = 12$. Here, systems up to size $N = 140$ are investigated. Since data for several different graph sizes are available, it is possible to extrapolate the behavior of the infinite graph using finite-size scaling techniques. The results of this extrapolations will be presented in a subsequent chapter, along with the outcomes of analytical calculations.

In Fig. 1 the probability $P_{cov}(x)$ of finding a vertex cover of cardinality xN for a random graph $G_{N,cN}$ is displayed for $c = 1$ and different values of N (10000 instances per value of x, 1000 for $N = 100$). The drop of the probability from one for large cover sets to zero for small cover sets obviously sharpens with N. Thus, a jump at a well-defined $x_c(c)$ is to be expected in the large-N limit: Above $x_c(c)$ almost all random graphs with cN edges are coverable with xN vertices, below $x_c(c)$ almost no graph has such a VC. The curves in the left part of the figure show the average minimal fraction $e(x)$ of uncovered edges, which for a coverable graph is obviously zero. In the large-N limit, the disappearance of positive $e(x)$ coincides with the threshold.

It is very instructive to measure the median computational effort, as given by the number of visited nodes in the backtracking tree, in dependence on x and N. The curves which are exposed in Fig. 2, show a pronounced peak at the threshold value. Inside the coverable phase, $x > x_c(c)$, the computational cost is growing only linearly with N, and in many cases the heuristic is already able to find a cover with xN vertices. Below the threshold, $x < x_c(c)$, it is clearly exponential in N (see inset). This easy-hard transition resembles very much the typical-case complexity pattern of 3SAT [19], and deserves some more detailed investigation, which will be provided by the analytical calculation later on.

In Fig. 3 the median time is plotted separately for the subset of coverable and uncoverable graphs, respectively. In addition, a scatter plot is included, which contains a dot for each result for 100 graphs and for different cardinalities xN of the cover. For a given graph $G_{N,cN}$, as long as it is not coverable with xN vertices, the computer time grows heavily with x. But as soon as a graph is coverable, it takes only a small computational effort to find a cover. The reason that median effort over all graphs is reduced for $x > x_c$ is that the fraction of uncoverable graphs decreases rapidly.

Another quantity is directly related to the transition: The outcome of the algorithm is a configuration, *i.e.* a vector of marks telling whether a given vertex is covered or not. For a given graph and a given fraction x usually different configurations are feasible, exhibiting all the same minimal number $e(x)cN$ of uncovered edges. An enumeration shows that the number of these configurations grows exponentially with the system size for all values of N. Nevertheless, for $x < x_c(c)$ there is always a finite fraction of vertices which behave equally in all different configurations: they are either always covered or always uncovered. The set of these vertices is the backbone B.

For $x > x_c$ and in the large-N limit, there is no non-empty backbone: the graph is already coverable with $x_c(G_{N,cN})N$ vertices, the other $(x - x_c)N$ can be distributed freely. This already excludes the existence of vertices being always uncovered. The maximal vertex degree in a random graph $G_{N,cN}$ grows only as $O(\log N)$. So the neighbors of every covered vertex can be covered with some of the remaining $(x-x_c)N$ free cover marks, and the central vertex itself can be uncovered and thus does not belong to the backbone.

Later we will see that directly at the threshold $x = x_c$ a finite backbone size $b(x) = |B|/N$ appears. Thus, for $N \to \infty$ the function $b(x)$ exhibits a discontinuity at $x_c(c)$. This is indicated by the results obtained from the numerical calculations, again for the case $c = 1$, see Fig. 4. For $x < x_c(1)$ the relative backbone size $b(x)$ is large and almost independent of N. For $x > x_c(1)$ a sharp decrease can be observed, which pronounces with increasing N. A surprising result is obtained, when we study coverable and uncoverable graphs separately. This can be done only in the vicinity of the transition, $x \approx x_c(1)$, where coexisting coverable and uncoverable graphs can be found for finite N. The inset of Fig. 4 shows the result: Above the threshold, the coverable graphs exhibit a smaller backbone, as expected from the discussion above. But the curves intersect near $x_c(1)$. This behavior is observed for all graph sizes N, and the effect becomes more pronounced with increasing system size. As an explanation, we take a look at graphs being coverable with a small number of vertices. Their distribution of vertex degrees must deviate substantially from (3), showing more vertices with high degree. These vertices are expected to be in the backbone with high probability, see also the discussion on the correlation between vertex degree and backbone at the end of section 5.3.1. Consequently, the backbone is expected to be very large. The crossing of both curves close to x_c seems to be accidental. By measuring the intersection as a function of N and extrapolating to $N \to \infty$, the limiting value is found to be significantly below x_c.

We have seen that the vertex-cover problem exhibits several peculiar features. These are worth to be addressed by analytical methods which allow to reveal the structure of VCs.

4 Cluster expansion for low vertex degrees

One of the classical results on random graphs is, as mentioned in section 2.2, that for low edge densities $c < 1/2$ almost all vertices are collected in finite trees, as

$$1 = \sum_{k=1}^{\infty} \rho(k) k^{k-2} k \tag{10}$$

with $\rho(k)$ being the distribution of trees T_k with k vertices, cf. section 2.2. So the threshold x_c and the corresponding backbone b are given by

$$\begin{aligned} x_c(c) &= \sum_{k=1}^{\infty} \rho(k) \left[\sum_{T_k} X_c(T_k) \right] \\ b(c, x_c(c)) &= \sum_{k=1}^{\infty} \rho(k) \left[\sum_{T_k} B_c(T_k) \right] \end{aligned} \tag{11}$$

where $\sum_{T_k}$ denotes the sum over all different trees T_k. $X_c(T_k)$ (resp. $B_c(T_k)$) is the cardinality of the minimal VCs (resp. of their backbone) of T_k.

For very small average vertex degrees $c \ll 0.5$ the most vertices are furthermore concentrated in small components, and we can produce good approximations for the threshold, the backbone etc. by counting small trees. There also the distinction between backbone and non-backbone vertices becomes evident: Consider e.g. a connected component consisting only of two vertices and one edge. To cover this minimally, we need exactly one vertex – but it is not specified which one. The vertices do not belong to the backbone at threshold, and they give a contribution to a finite entropy (*i.e.* an exponential number) of minimal VCs. The situation is different for a tree of three vertices and two edges. The minimal cover is unique: Only the central vertex has to be taken. Consequently all these three vertices belong to the backbone at the threshold. Already at this point, the partial freezing of degrees of freedom as observed in SAT [17, 19] becomes evident.

We have counted the optimal covers for trees up to 7 vertices, see the results in table 1. The values for the threshold and the backbone are lower bounds as a certain fraction of vertices is not included. Upper bounds are provided by adding the fraction of missing vertices to the lower bounds. For small c these bound are very precise, *e.g.* for $c = 0.1$, 99.98% of all vertices are already included in the small trees up to size 7. These approximate values will be a useful testing ground for the statistical mechanics calculations which are given in section 5.

This tree size expansion is not longer possible above the percolation threshold $c = 1/2$. There the giant component arises which includes a finite fraction of all vertices.

5 Statistical mechanics approach

In this section we use the strong similarities between combinatorial optimization and statistical mechanics. The cost function of a system which shall be optimized corresponds to the *energy function* (or *Hamiltonian*) in statistical mechanics. The elements of the definition space of the cost function are called microscopic *configurations.* The main aim of statistical mechanics is the description of the macroscopic behavior of a microscopically defined model, *e.g.* the prediction and description of phase transitions.

c	0.05	0.1	0.15	0.2	0.25	0.3	0.4	0.5
ν	0.999997	0.9998	0.998	0.991	0.97	0.94	0.84	0.71
x_{min}	0.045576	0.0840	0.116	0.143	0.16	0.17	0.17	0.15
x_{max}	0.045579	0.0842	0.118	0.151	0.19	0.23	0.33	0.44
b_{min}	0.916684	0.8572	0.812	0.774	0.74	0.70	0.61	0.51
b_{max}	0.916687	0.8574	0.814	0.781	0.77	0.76	0.77	0.80
s_{min}	0.028774	0.0488	0.063	0.073	0.078	0.08	0.08	0.07
s_{max}	0.028775	0.0489	0.064	0.076	0.088	0.10	0.13	0.17
$x_c(c)$	0.045577	0.0841	0.117	0.146	0.173	0.196	0.237	0.272
$b_c(c)$	0.916686	0.8573	0.813	0.779	0.753	0.731	0.700	0.678

Table 1: Results of the cluster expansion for trees having up to 7 vertices and several values of c. ν denotes the fraction of vertices which are included in the considered trees, $Nx_{min/max}$ give lower and upper bounds on the number of vertices which are needed to cover these components, $b_{min/max}$ are backbone bounds, $s_{min/max}$ bounds for the VC entropy. These values are to be compared with the analytical results of the replica approach which are presented in the last two lines.

5.1 General strategy

In order to describe the VC phase transition also beyond the percolation threshold, we are going to use the tools of the statistical mechanics of disordered systems [15]. We therefore map the random graph to a disordered spin system with an Hamiltonian which shall be minimized. A canonical choice for the "energy" of a subset $\tilde{V} \subset V$ of vertices is given by the number of uncovered edges:

$$H(\{S_i\},\{J_{i,j}\}) = \frac{1}{2}\sum_{i,j=1}^{N} J_{i,j}\delta_{S_i,-1}\delta_{S_j,-1} \tag{12}$$

where $J_{i,j}$ are the entries of the symmetric adjacency matrix, they are equal to one whenever there is an edge connecting the vertices i and j, and zero else. The diagonal elements are identically set to zero. The covering state of the vertices is mapped to a configuration of N Ising-spins $S_i = \pm 1$: we choose $S_i = +1$ if $i \in \tilde{V}$, *i.e.* if the vertex i is covered, and $S_i = -1$ if i is uncovered. Non-zero contributions to the Hamiltonian result only from edges having two uncovered endpoints.

The decision problem whether there exists any VC with xN vertices can be answered by minimizing H under the constraint

$$\frac{1}{N}\sum_{i=1}^{N} S_i = 2x - 1 \tag{13}$$

which fixes the cardinality of the cover set, or in physical terms, the global magnetization of our Ising-spin system. If this restricted minimal energy equals zero, then there are no uncovered edges left, and the decision problem can be positively answered. If, on the other hand, a positive minimal energy is found, there does not exist any VC of cardinality xN, but the ground state energy gives the best compromise by describing the configuration with the minimal number of uncovered edges.

In statistical mechanics every microscopic configuration $\{S_i\}_{i=1,..,N}$ is assigned a probability proportional to the Gibbs-weight $\exp\{-T^{-1}H(\{S_i\})\}$ at temperature T. By decreasing

T, this weight becomes more and more concentrated in low-energy configurations and finally, at $T = 0$, counts only the *ground states*, *i.e.* the configurations minimizing the Hamiltonian. In order to characterize these in the VC problem, we introduce at first a *non-zero formal temperature* T and calculate the *partition function*

$$Z(T, x|\{J_{i,j}\}) = \sum_{\mathcal{C}_x(\{S_i\})} \exp\left\{ -\frac{H(\{S_i\}, \{J_{i,j}\})}{T} \right\} \tag{14}$$

where we sum only over the set $\mathcal{C}_x(\{S_i\})$ of configurations $\{S_i\}_{i=1,..,N}$ which satisfy the magnetization constraint (13). From this we may calculate the *free-energy density*

$$f(T, x|\{J_{i,j}\}) = -\frac{T}{N} \log Z(T, x|\{J_{i,j}\}) \tag{15}$$

which in its zero temperature limit gives the desired *ground state energy density*:

$$e_{GS}(x|\{J_{i,j}\}) = \lim_{T\to 0} f(T, x|\{J_{i,j}\}) \ . \tag{16}$$

This energy does still depend on the particular realization of the graph encoded in the matrix $\{J_{i,j}\}$. In the limit $N \to \infty$ (with $c = M/N = const.$) we expect however the free energy to be *self-averaging* , and so we are only interested in calculating

$$e_{GS}(x, c) = \lim_{T\to 0} f(T, x, c) = \lim_{T\to 0} \lim_{N\to\infty} \overline{f(T, x|\{J_{i,j}\})} \tag{17}$$

where the over bar stands for the average over the ensemble of random graphs with N vertices and cN edges. Another interesting quantity is the *ground state entropy*

$$s_{GS}(x, c) = \lim_{N\to\infty} \frac{1}{N} \overline{\log \mathcal{N}_{GS}(x, \{J_{i,j}\})} \tag{18}$$

where $\mathcal{N}_{GS}(x, \{J_{i,j}\})$ is the number of ground states with cardinality xN in the graph given by $\{J_{i,j}\}$. It is also useful to consider the *VC entropy*

$$s_{VC}(x, c) = \begin{cases} s_{GS}(x, c) & \text{if } e_{GS}(x, c) = 0 \\ -\infty & \text{else} \end{cases} \tag{19}$$

which measures the number of VCs.

5.2 The annealed approximation

Before trying to calculate this, we will present the so-called *annealed approximation.* We use the bound

$$\overline{\log Z(T, x|\{J_{i,j}\})} \le \log \overline{Z(T, x|\{J_{i,j}\})} \tag{20}$$

for the average of the logarithm of the partition function in terms of the logarithm of the average of the partition function. It holds because the logarithm is a concave function. We easily calculate the annealed entropy, see Appendix A for details,

$$\begin{aligned} s_{ann}(x, c) &= \lim_{T\to 0} \lim_{N\to\infty} \frac{1}{N} \log \overline{Z(T, x|\{J_{i,j}\})} \\ &= -x \log x - (1 - x) \log(1 - x) - c(1 - x)^2 \end{aligned} \tag{21}$$

and can bound the VC entropy

$$s_{VC}(x, c) \le s_{ann}(x, c) \ . \tag{22}$$

VCs can thus only exist if the annealed entropy is non-negative, and $x_c(c)$ is bounded from below by $x_{ann}(c)$ which is given by $s_{ann}(x_{ann}(c),c)=0$, *i.e.* by the inversion of

$$c=\frac{-x_{ann}(c)\log x_{ann}(c)-(1-x_{ann}(c))\log(1-x_{ann}(c))}{(1-x_{ann}(c))^2}\ . \tag{23}$$

This is exactly the lower bound given in [7] which is not surprising as Gazmuri used a very similar reasoning.

5.3 The replica approach

If we want to go beyond the annealed approximation, we have to average the logarithm of the partition function over the disorder. Unfortunately this cannot be achieved directly, the way out is given by the so-called *replica trick*, a non-rigorous method which is well-established in the physics of disordered systems [15]. Details of the calculation are exposed in appendix B. There we show the derivation of the so-called *replica symmetric approximation* of the free-energy density

$$\begin{aligned} f(T,x,c) &= T\int_{-\infty}^{\infty}\frac{dh\ dk}{2\pi}e^{-ihk}P_{FT}(k)[\log P_{FT}(k)-1]\log 2\cosh T^{-1}h \qquad (24)\\ &\quad -cT\int_{-\infty}^{\infty}dh_1dh_2P(h_1)P(h_2)\log\left[1-(e^{-T^{-1}}-1)\frac{e^{-T^{-1}(h_1+h_2)}}{4\cosh T^{-1}h_1\ \cosh T^{-1}h_2}\right]\ . \end{aligned}$$

This quantity has to be optimized with respect to the order parameter $P(h)$ which is again restricted by the magnetization constraint to

$$2x-1=\int_{-\infty}^{\infty}dhP(h)\tanh T^{-1}h\ . \tag{25}$$

$P_{FT}(k)$ denotes the Fourier-transform of $P(h)$.

The physical interpretation of the order parameter in terms of the *effective field distribution* is straightforward: $P(h)dh$ gives the probability, that a randomly chosen site $i\in V$ has local magnetization $m_i=\langle S_i\rangle_T=\tanh T^{-1}h$. This distribution (or the distribution of local magnetizations) is the typical order parameter in disordered finite connectivity models, cf. [11, 18]. It is determined by the optimization equation for the free energy (24) which reads

$$\int dh\ P(h)\ e^{T^{-1}hs}=\exp\left\{-2c-\lambda s+2c\int dh\ P(h)\ \left[1+(e^{-T^{-1}}-1)\frac{1}{1+e^{2T^{-1}h}}\right]^{-\frac{s}{2}}\right\}\ . \tag{26}$$

The Lagrange parameter λ in the exponential has to be adjusted in order to meet the magnetization constraint (25).

This equation as well as the expression (24) for the free energy still depend on the formal temperature T, and the limit $T\to 0$ is not totally obvious: we have to clarify the scaling of the effective fields h with T. There are two main possibilities:

- The fields h are proportional to the formal temperature, $h=O(T)$ for $T\to 0$. As can be simply seen in the expression (24) for the average free energy, we then also have $f(T,x,c)=O(T)$, and the ground state energy $e_{GS}(x,c)$ vanishes. These fields are consequently found in the coverable phase with $x>x_c$. Another important property is that the corresponding local magnetizations $m=\tanh T^{-1}h$ do not tend to ±1, and the corresponding spins take different orientations in different ground states.

- The fields h remain different from 0 even if the temperature vanishes, $h = O(T^0)$. The corresponding spins have ±1-ground state magnetization, and consequently take on the same value in (almost) all ground state configurations, *i.e.* they form the backbone. If we introduce such fields in (24) we immediately find that $f(T, x, c)$ does not vanish in the zero-temperature limit, the ground state energy becomes positive, and such fields cannot exist in the COV phase. Their appearance marks the transition.

5.3.1 At the threshold

If we would be able to solve (26) at finite temperature for arbitrary x and c, we could deduce the scaling directly from the solution – and thus we could determine $x_c(c)$. As this is to complicated to be achieved directly, we can plug in the two different scalings, and calculate the limit $T \to 0$. We then find two different equations for $P(h)$ in the two different phases. The phase transition point is given by the matching of both equations:

- If we reach the threshold from above, $x \to x_c(c) + 0$, we are in the coverable phase. According to the above discussion, the effective fields are $h = TH_{cov}(x)z$ where $H_{cov}(x)$ describes the typical absolute value of the field and z is a random variable of finite mean and variance. For $x \to x_c(c) + 0$ the spins S_i are more and more constraint, and at $x_c(c)$ a freezing takes place. The limit is therefore described by $H_{cov}(x \to x_c(c) + 0) \to \infty$.
- If we reach the threshold from below, $x \to x_c(c) - 0$, we are in the uncoverable phase, and at least a finite fraction of all spins has to be frozen. The corresponding effective fields scale as $h = H_{uncov}(x)z$ where now the scale for the absolute value of h is described by $H_{uncov}(x)$. As we approach the threshold, the freezing gets less strong, and $H_{uncov}(x \to x_c(c) + 0) \to 0$.

In both limits we find the same equation for the probability distribution $\tilde{P}(z)$ of the rescaled variable z, see appendix B for a derivation. μ is the appropriately rescaled Lagrange parameter, it is negative as it describes a field which decreases the global magnetization from the maximum entropy point towards the threshold $x_c(c)$:

$$\begin{aligned} \tilde{P}(z+\mu) &= \sum_{d=0}^{\infty} e^{-2c}\frac{(2c)^d}{d!}\left[\tilde{P_-}^{\star d}\right](z) \\ \tilde{P_-}(z) &= \Theta(z)\tilde{P}(-z) + \delta(z)\int_{-0}^{\infty} dz\ \tilde{P}(z) \end{aligned} \tag{27}$$

with the Heaviside step function $\Theta(z)$ and the Dirac distribution $\delta(z)$. $\tilde{P_-}^{\star d}$ denotes a d-fold convolution product. The interpretation of this equation is simple: the effective field for a randomly chosen vertex i is given by the linear superposition of the local field induced by the Lagrangian multiplier, and the contribution of its d_i neighbors. If a neighbor has a negative field, then it is uncovered, and thus forces a positive field on i. If it has a non-negative field it does not imply any non-vanishing field on i. As $\tilde{P}(z)$ is the histogram of fields for all vertices, equation (27) includes the average over the Poisson distribution (3) of vertex degrees.

This equation has a very simple solution,

$$\tilde{P}(z) = \sum_{m=-1}^{\infty} \frac{W(2c)^{m+2}}{2c(m+1)!}\delta(z+m\mu)\ , \tag{28}$$

with the Lambert-W function W which is simply defined by

$$y = W(x) \leftrightarrow x = y\ e^y\ . \tag{29}$$

Non-zero fields correspond to frozen (or backbone) spins, whereas the Dirac peak in $z = 0$ describes all spins which flip from one minimal VC to a next. The backbone size is consequently given by the total weight of all nonzero fields. From this we can calculate the threshold and the backbone,

$$\begin{aligned} x_c(c) &= 1 - \frac{2W(2c) + W(2c)^2}{4c} \\ b_c(c) &= 1 - \frac{W(2c)^2}{2c} \ . \end{aligned} \tag{30}$$

This result is completely consistent with the bounds of section 4 which is particularly interesting for very small c where these bounds are very close, see table 1. The result for $x_c(c)$ is displayed in Fig. 5 along with numerical data, which were obtained by the variant of the branch-and-bound algorithm which always looks for a cover of minimum cardinality. For each treated concentration c of the edges and system sizes $N = 12$, 17, 25, 35, 50, 70, 100, 140 for 10000 different realizations of the random graphs (only 1000 for the $n \geq 100$) the threshold was calculated. The average value is denoted with $x_c(c, N)$. Then for each value of c the behavior of the infinite graph was extrapolated by performing a fit of the function $x_c(N) = x_c + aN^{-b}$ to the data, where x_c, a and b are tunable parameters. The inset shows an example of such a kind of extrapolation. The result of x_c as a function of c shows a very good coincidence with the analytic result. This is true not only for small concentrations but also for a region beyond the percolation threshold, whereas systematic deviations appear for larger c.

Are there more complicated solutions to (27) which coincide with the numerics also for larger c? At first we remark that this equation is closed under

$$\tilde{P}^{(l)}(z) = \sum_{m=-l}^{\infty} a_m^{(l)} \delta\left(z - m\frac{\mu}{l}\right) \tag{31}$$

for every positive integer l. The equations for $a_{-l}^{(l)}, ..., a_{-1}^{(l)}$ close, all other weights with non-negative indices follow. A simple analysis of these equations shows, that for $c < e/2$ they have no non-trivial solution with only non-negative weights, up to this point (28) gives the only valuable solution. For $c > e/2$ non-trivial solutions with an arbitrary number of peaks appear.

Together with the above mentioned accordance of bounds and numerical data for low vertex degrees, this leads to the following **conjecture**: *For random graphs with $c \leq e/2$ the exact values for the covering threshold and the backbone at this threshold are given by equation (30). For $c > e/2$, the above value for $x_c(c)$ still gives a lower bound.*

The last statement follows from the fact that in the replica approach the saddle point with the largest free energy has to be taken. Imagine now two different values for x_c would be predicted by two different saddle points. In between these thresholds, one solution already predicts a positive energy and hence a larger free energy than the other. This saddle point has to be preferred, and it corresponds to the larger threshold.

The transition at $c = e/2$ is not yet understood as also the multi-peak solutions (31) do not coincide with numerical data. This can be seen in particular from the behavior of the backbone size – which is largely overestimated analytically, see Fig. 6. Especially the minimum of $b_c(c)$ at $c = e/2$ cannot be found in the numerical data. The numerical results were obtained from the enumerating of all possible covers at the threshold for the same range of concentrations and sizes mentioned above. Also the same extrapolation technique to obtain the values for the infinite random graph was applied.

For the discrepancy of the numerical backbone size with the analytical data for $c > e/2$, there are two possible explanations:

- In the analytics we count every spin as backbone which has magnetization tending to ±1 in the thermodynamic limit, whereas the in numerics we count only vertices which have magnetization equal to ±1 even for finite size. This difference can be rather drastic: *e.g.* for the fully connected graph of N vertices one needs $N-1$ for a VC, the average magnetization is therefore $1-2/N$. The analytics would count a backbone one, whereas the strict backbone vanishes.
- Above $c = e/2$ (or even above $c = 1/2$) replica symmetry breaking could appear. This would correspond to a clustering of the VCs in configurations space, cf. [1] for a discussion of this phenomenon for SAT. As was seen there, the backbone size sensitively depends on this question. This point is still under investigation.

Let us go back to $0 < c < e/2$ where (28) was conjectured to be exact, and let us extract more information about the minimal VCs from our solution. Due to the simple geometrical nature of the underlying graphs, the VC problem allows a much more intuitive way of understanding results, in contrast for example to SAT. A first example was already given in section 4 where we gave simple examples for backbone and non-backbone structures. Let us now investigate the influence of the close environment of a vertex on its behavior, more precisely the influence of the vertex degree. The total distribution of (almost all) degrees is given by the Poisson law (3), but we can distinguish three distinct contributions:

- The joint probability $P(d, m = -1)$ that a vertex has degree d and magnetization $m = -1$, *i.e.* this vertex belongs to the backbone and is uncovered in all minimal VCs.
- $P(d, m = +1)$ gives the probability that a vertex has degree d and is covered in all minimal VCs.
- The remaining part of vertices are not in the backbone, thus described by $P(d, -1 < m < +1)$.

These quantities can be easily computed from $\tilde{P}(z)$: according to the interpretation of the self-consistent equation (27) we can calculate the effective-field distribution for a vertex of degree d which, in average, has typical neighbors:

$$\tilde{P}_d(z + \mu) = \left[\tilde{P_-}^{\star d}\right](z) \tag{32}$$

where $\tilde{P_-}(z)$ is exactly the quantity given in (27). Plugging our solution (28) into this equation, we find

$$\begin{aligned} P(d, m = -1) = \tilde{P}_d(z < 0) Po_{2c}(d) &= e^{-2c} \frac{[2c - W(2c)]^d}{d!} \\ P(d, -1 < m < +1) = \tilde{P}_d(z = 0) Po_{2c}(d) &= e^{-2c} \frac{W(2c)[2c - W(2c)]^{d-1}}{(d-1)!} \\ P(d, m = +1) = \tilde{P}_d(z > 0) Po_{2c}(d) &= e^{-2c} \frac{[2c + (d-1)W(2c)][2c - W(2c)]^{d-1}}{(d-1)!} \end{aligned} \tag{33}$$

The results for $c = 1$ are displayed in Fig. 7 along with numerical data for $N = 17, 35, 70$. Please note that the numerical results seem to converge towards the analytical one, thus showing an excellent coincidence of both approaches. The curves are easily understood: a vertex with degree 0 has no neighbors. Therefore, it does not appear in any optimum cover and we obtain $P(0, m = -1) = 1$, $P(0, m > -1) = 0$. With increasing degree the probability that a vertex is covered increases, thus the contribution of $P(k, -1 < m < +1)$ to $\mathrm{Po}_{2c}(d)$

increases as well. For large degrees it is very probable that a vertex belongs to all VCs but even a finite fraction of vertices with $m = -1$ remains.

This behavior can also be studied by evaluating the average magnetization $m(d)$ as a function of the degree. Here the analytical solution gives only lower and upper bounds since we are not able to precisely calculate the magnetization of the non-backbone spins:

$$2\left(1 + (d-1)\frac{W(2c)}{2c}\right)\left(1 - \frac{W(2c)}{2c}\right)^{d-1} - 1 < m(d) < 1 - 2\left(1 - \frac{W(2c)}{2c}\right)^{d} \tag{34}$$

Results are displayed in Fig. 8: with increasing size N of the graphs the numerical data approach the region inside the bounds. The magnetization turns out to be a monotonously increasing function of the vertex degree, as expected from the results for $P(k,m)$. These results justify *a posteriori* the application of the heuristic within the algorithm: vertices having a large degree are at first included into the cover set.

5.3.2 Approximating the VC entropy

It is also interesting to go away from the threshold into the coverable phase, $x > x_c(c)$, and to ask for the number of VCs which is given by the cover entropy (19). As the saddle point equations for $P(h)$ are to hard to be solved directly, we have used a simple variational ansatz. For doing this, we plug a set of simple test functions into the free energy (24) and optimize with respect to these, cf. [1] for an application in SAT. The simplest Ansatz is provided by taking a Gaussian distribution,

$$P^{(var)}(h) = \frac{1}{\sqrt{2\pi\Delta}T}\exp\left\{-\frac{(h-Tz_0)^2}{2\Delta T^2}\right\} \tag{35}$$

which includes only two free parameters. Note that the resulting fields h have already the linear scaling with temperature T which is needed for the limit $T \to 0$ in the coverable phase. Using the rescaled variable $z = h/(T\sqrt{\Delta})$, we get the following variational expression for the VC entropy:

$$\begin{aligned} s_{VC}^{(var)}(x,c) &= \int Dz \frac{3 - z(z - 2z_0/\sqrt{\Delta})}{2}\log[2\cosh(\sqrt{\Delta}z + z_0)] \\ &\quad + c\int Dz_1 \int Dz_2 \log\left[1 - \frac{\exp\{-\sqrt{\Delta}(z_1+z_2) - 2z_0\}}{4\cosh(\sqrt{\Delta}z_1 + z_0)\cosh(\sqrt{\Delta}z_2 + z_0)}\right] \end{aligned} \tag{36}$$

Dz denotes the normal Gaussian measure $dz\ e^{-z^2/2}/\sqrt{2\pi}$. This expression has to be optimized with respect to the parameters Δ and z_0 which fulfill the additional constraint

$$\int_{-\infty}^{\infty} Dz \tanh(\sqrt{\Delta}z + z_0) = 2x - 1\ . \tag{37}$$

Fig. 9 compares the resulting entropy with numerical enumerations of all VCs for graphs with $c = 1.0$ as a function of x. Because of the large numerical effort, only graphs with $N \leq 50$ were considered. Deep inside the coverable region, the value of $s_{VC}^{(var)}$ appears to be a very good approximation, as the numerical values approach it with increasing graph sizes N. Near the threshold the Gaussian ansatz (35) starts to fail as it includes only one scale for the fields and thus is not able to reflect the partial freezing into backbone and non-backbone spins. Comparable results were also obtained for other values of c.

6 Conclusions and outlook

In this paper the vertex-cover problem on random graphs with a finite average vertex degree was studied. The problem was investigated using several methods. Numerical calculations with an exact branch-and-bound algorithm were performed. The coverability of a graph shows a sharp transition in the cardinality xN of vertex covers at the threshold $x_c(c)$. There are almost surely no VCs with $x < x_c(c)$, whereas they exist almost surely for $x > x_c(c)$. This transitions is related to a jump in the median complexity of the algorithm, and in the size of the backbone as well.

A cluster expansion for non-percolated graphs gives very precise estimates of threshold and backbone for small c. Two approaches coming from the statistical physics of disordered systems were applied to the VC problem. The annealed approximation reproduces a known graph-theoretical lower bound. A more sophisticated method is given by the replica ansatz, which allows to derive analytical expression for the threshold $x_c(c)$ and the backbone $b_c(c)$ for average vertex degrees less than the Eulerian constant e, where also the agreement with numerical data is excellent. These expressions are conjectured to be exact. Beyond the average connectivity $2c = e$, the replica symmetric ansatz fails to produce valuable results, and more complicated methods including replica symmetry breaking should be applied in future.

We have also given a variational approximation for the vertex cover entropy, *i.e.* the logarithm of the number of VCs of given cardinality. Whereas this approximation was rather precise far above the covering threshold, the latter can be described only by going beyond a simple Gaussian approximation. The behavior for $x \neq x_c(c)$ deserves further investigation.

It would also be interesting to consider different graph ensemble, *e.g.* graphs of constant vertex degree or graphs having locally non-tree-like structures.

7 Acknowledgements

The authors are deeply indebted to R. Monasson and R. Zecchina for many fruitful discussions on the field of phase transitions in combinatorial problems. Financial support was provided by the DFG (*Deutsche Forschungsgemeinschaft*) under grant Zi209/6-1.

A Calculation of the annealed bound

In this appendix we calculate the annealed bound for the covering threshold. As stated in (20), it follows from the average of the partition function over the random graph ensemble. Here we use the second formulation, see 2.2, where edges are drawn with probability $2c/N$:

$$\begin{aligned}
\overline{Z(T,x|\{J_{i,j}\})} &= \sum_{\mathcal{C}_x(\{S_i\})} \overline{\exp\{-H(\{S_i\},\{J_{i,j}\})/T\}} \\
&= \sum_{\mathcal{C}_x(\{S_i\})} \prod_{1\le i<j\le N} \overline{\exp\{-J_{i,j}\delta_{S_i,-1}\delta_{S_j,-1}/T\}} \\
&= \sum_{\mathcal{C}_x(\{S_i\})} \prod_{1\le i<j\le N} \left[1 - \frac{2c}{N} + \frac{2c}{N}\exp\{-\delta_{S_i,-1}\delta_{S_j,-1}/T\}\right] \\
&= \sum_{\mathcal{C}_x(\{S_i\})} \exp\left\{-cN + \frac{c}{N}\sum_{i,j=1}^{N}\exp\{-\delta_{S_i,-1}\delta_{S_j,-1}/T\} + o(N)\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \binom{N}{xN} \exp\left\{ cN(1-x)^2(e^{-T^{-1}}-1) + o(N) \right\} \\
&= \exp\left\{ N\left[-x\log(x) - (1-x)\log(1-x) + c(1-x)^2(e^{-T^{-1}}-1) \right] + o(N) \right\}
\end{aligned}
$$

where the last expression was obtained using Stirlings formula. This gives the annealed entropy from section 5.2 in the limit $T \to 0$.

B Calculation of the free energy

The main problem in calculating the free-energy density consists in the average of the logarithm of the partition function over the ensemble of random graphs. The replica trick is based on the simple equality

$$
\log Z = \lim_{n\to 0} \frac{Z^n - 1}{n} \tag{38}
$$

which is valid for positive real Z. It allows to calculate the average of Z^n. In principle, this problem is not easier than before. But the trick used in statistical physics is the following: We calculate $\overline{Z^n}$ at first for positive integer n, and try to obtain some analytical continuation at the end. The n-fold power can be understood in terms of n identical copies $\{S_i^a\},\ a = 1, ..., n$, of the original system. Every of these copies has the same Hamiltonian (12), including identical edges $J_{i,j}$, and fulfills the same magnetization constraint (13). The average over random graphs is calculated analogously to the last appendix, cf. section (5.1) for the notations,

$$
\begin{aligned}
Z_n(T,x,c) &:= \overline{Z^n(T,x|\{J_{i,j}\})} \\
&= \sum_{\mathcal{C}_x(\{S_i^a\})} \overline{\exp\left\{ -T^{-1} \sum_{i<j} J_{i,j} \sum_{a=1}^{n} \delta_{S_i^a,-1}\delta_{S_j^a,-1} \right\}} \qquad (39) \\
&= \sum_{\mathcal{C}_x(\{S_i^a\})} \exp\left\{ -cN + \frac{2c}{N} \sum_{i<j} \exp\left\{ -T^{-1} \sum_{a=1}^{n} \delta_{S_i^a,-1}\delta_{S_j^a,-1} \right\} + o(N) \right\}
\end{aligned}
$$

This can be simplified by introducing the 2^n order parameters which are enumerated by $\vec{\sigma} \in \{+1,-1\}^n$:

$$
c(\vec{\sigma}) = \frac{1}{N} \sum_{i=1}^{N} \prod_{a=1}^{N} \delta_{\sigma^a, S_i^a} \ . \tag{40}
$$

$c(\vec{\sigma})$ measures the fraction of vertices i having the replicated spin $(S_i^1, ..., S_i^n) = \vec{\sigma}$. We find

$$
\begin{aligned}
Z_n(T,x,c) &= \int_0^1 \prod_{\vec{\sigma}}{}' dc(\vec{\sigma}) \frac{N!}{\prod_{\vec{\sigma}}[c(\vec{\sigma})N]!} \times \qquad (41) \\
&\quad \times \exp\left\{ -cN + cN \sum_{\vec{\sigma},\vec{\tau}} c(\vec{\sigma})c(\vec{\tau}) \exp\left\{ -T^{-1} \sum_{a=1}^{n} \delta_{\sigma^a,-1}\delta_{\tau^a,-1} \right\} + o(N) \right\}
\end{aligned}
$$

The integration is over all $c(\vec{\sigma})$ which are normalized, $\sum_{\vec{\sigma}} c(\vec{\sigma}) = 1$, and fulfill the magnetization constraint, $\sum_{\vec{\sigma}} c(\vec{\sigma})\sigma^a = 2x - 1$ for all $a = 1, ..., n$. Using Sterlings formula we finally find

$$
Z_n(T,x,c) = \int_0^1 \prod_{\vec{\sigma}}{}' dc(\vec{\sigma}) \exp\left\{ N \left[-c - \sum_{\vec{\sigma}} c(\vec{\sigma}) \log c(\vec{\sigma}) \right.\right.
$$

$$
\begin{aligned}
& \left. \left. +c \sum_{\vec{\sigma},\vec{\tau}} c(\vec{\sigma}) c(\vec{\tau}) \exp \left\{ -T^{-1} \sum_{a=1}^{n} \delta_{\sigma^a,-1} \delta_{\tau^a,-1} \right\} + o(N) \right] \right\} \\
= \;& \exp \left\{ N g_n \left[c_0(\vec{\sigma}) \right] + o(N) \right\} \qquad (42)
\end{aligned}
$$

The dominant term of $O(N)$ in the exponent is given by the saddle point $c_0(\vec{\sigma})$,

$$
\log c_0(\vec{\sigma}) = \lambda_1 + \lambda_2 \sum_a \sigma^a + 2c \sum_{\vec{\tau}} c_0(\vec{\tau}) \exp \left\{ -T^{-1} \sum_{a=1}^{n} \delta_{\sigma^a,-1} \delta_{\tau^a,-1} \right\} \qquad (43)
$$

where λ_1 is a Lagrange parameter for the normalization of $c(\vec{\sigma})$, and λ_2 a second one for the magnetization constraint.

The problem which remains is the continuation to real n. We have to introduce some ansatz on the structure of $c_0(\vec{\sigma})$. The simplest one is based on the observation, that $Z_n(T,x,c)$ is by definition invariant under permutations of the n replicas which were introduced as being identical. We therefore assume this symmetry also for the order parameter $c_0(\vec{\sigma})$ which consequently depends only on $s = \sum_a \sigma^a$. We may express it by a generating function,

$$
c_0(\sigma) = \int_{-\infty}^{\infty} dh\ P(h) \frac{e^{T^{-1} h s}}{(2 \cosh T^{-1} h)^n}, \qquad (44)
$$

which is normalized whenever $P(h)$ is normalized, $\int_{-\infty}^{\infty} dh\ P(h) = 1$. The magnetization condition now reads $\int_{-\infty}^{\infty} dh\ P(h)\ \tanh T^{-1} h = 2x - 1$.

Plugging this replica symmetric ansatz into $g_n[c_0(\vec{\sigma})]$, we get (24) by some straightforward algebra from

$$
f(T,x,c) = -T \lim_{n \to 0} \frac{1}{n} g_n[c_0(\vec{\sigma})] \ . \qquad (45)
$$

Also the saddle point equation (26) for $P(h)$ can be easily calculated from (43).

C The saddle point equation at the threshold

In order to calculate the saddle point equation at the threshold, we take the first procedure proposed in section 5.3.1, *i.e.* we approach the threshold from above, using the scaling $h = T H_{cov} z$ with some random variable z drawn from the distribution $\tilde{P}(z)$. In the limit $T \to 0$, (26) slightly simplifies ($\lambda = H_{cov}\mu$):

$$
\int dz\ \tilde{P}(z)\ e^{H_{cov} z s} = \exp \left\{ -2c - H_{cov} \mu s + 2c \int dz\ \tilde{P}(z) \ \left[\frac{1}{1 + e^{-2 H_{cov} z}} \right]^{-\frac{s}{2}} \right\} \ . \qquad (46)
$$

If we approach the threshold, H_{cov} is diverging. In order to obtain a reasonable limit, we have to keep $t = H_{cov} s$ finite in this limit:

$$
\begin{aligned}
\int dz\ \tilde{P}(z)\ e^{zt} \;&=\; \exp \left\{ -2c - \mu t + 2c \int dz\ \tilde{P}(z) \lim_{H_{cov} \to \infty} \left[\frac{1}{1 + e^{-2 H_{cov} z}} \right]^{-\frac{t}{2H_{cov}}} \right\} \\
&=\; \exp \left\{ -2c - \mu t + 2c \int_{-0}^{\infty} dz\ \tilde{P}(z) + 2c \int_{-\infty}^{-0} dz\ \tilde{P}(z) e^{-tz} \right\} \ . \qquad (47)
\end{aligned}
$$

Developing the exponential for the last two terms, we find the desired equation.

References

[1] G. Biroli, R. Monasson, and M. Weigt, to appear in *Europ. Phys. J.* **B**

[2] B. Bollobás and P. Erdös, *Math. Proc. Cambridge Philos. Soc.* **80**, 419 (1976)

[3] C. Caro, *Technical Report*, Tel Aviv University (1979)
V. K. Wei, *Bell Lab. Technical Memorandum* 81-11217-9 (1981)

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

[5] A. M. Frieze, *Discr. Math.* **81**, 171 (1990)

[6] M. R. Garey and D. S. Johnson, *Computers and Intractability*, (Freeman, San Francisco, 1979)

[7] P. G. Gazmuri, *Networks* **14**, 367 (1984)

[8] I. P. Gent and T. Walsh, *KI 94: Adv. Artif. Intell.* (Springer, Berlin 1994)

[9] J. Harant, *Discr. Math.* **188**, 239 (1998)

[10] B. Hayes, *Amer. Scient.* **85**, 108 (1997)

[11] I. Kanter and H. Sompolinsky, *Phys. Rev. Lett.* **58**, 164 (1987)

[12] E.L. Lawler and D.E. Wood, *Oper. Res.* **14**, 699 (1966)

[13] K. Mehlhorn and St. Näher, *The LEDA Platform of Combinatorial and Geometric Computing*, (Cambridge University Press, 1999);
see also http://www.mpi-sb.mpg.de/LEDA/leda.html

[14] S. Mertens, *Phys. Rev. Lett* **81**, 4281 (1998)

[15] M. Mézard, G. Parisi, and Virasoro, *Spin glasses and beyond*, (World Scientific, Singapore 1987)

[16] D. Mitchell, B. Selman, and H. Levesque, in *Proc. 10th Natl. Conf. Artif. Intell. (AAAI-92)*, 440 (AAAI Press/MIT Press, Cambridge, Massachusetts, 1992)

[17] R. Monasson, R. Zecchina, *Phys. Rev.* **E56**, 1357 (1997)

[18] R. Monasson, *J. Phys.* **A31**, 513 (1998)

[19] R. Monasson, R. Zecchina, S. Kirkpatrick, B. Selman, and L. Troyansky, *Nature* **400**, 133 (1999)

[20] T. Hogg, B. A. Huberman, and C. Williams (eds.), Frontiers in problem solving: phase transitions and complexity, *Artif. Intell.* **81** (I+II) (1996)

[21] R.E. Tarjan and A.E. Trojanowski, *SIAM J. Comp.* **6**, 537 (1977)

[22] M. Shindo and E. Tomita, *Syst. Comp. Jap.* **21**, 1 (1990)

[23] M. Weigt and A. K. Hartmann, subm. to *Phys. Rev. Lett.*, see also cond-mat/0001137

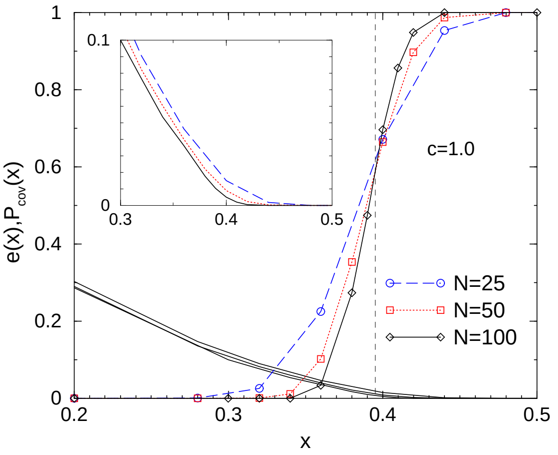

Figure 1: Probability $P_{cov}(x)$ that a cover exists for a random realization ($c = 1.0$) as a function of the fraction x of covered vertices. The result is shown for three different system sizes $N = 25, 50, 100$ (averaged for 10^4 - 10^3 samples). Lines are guides to the eyes only. In the left part, where the P_{cov} is zero, the energy e (see text) is displayed. The inset enlarges the result for the energy in the region $0.3 \leq x \leq 0.5$.

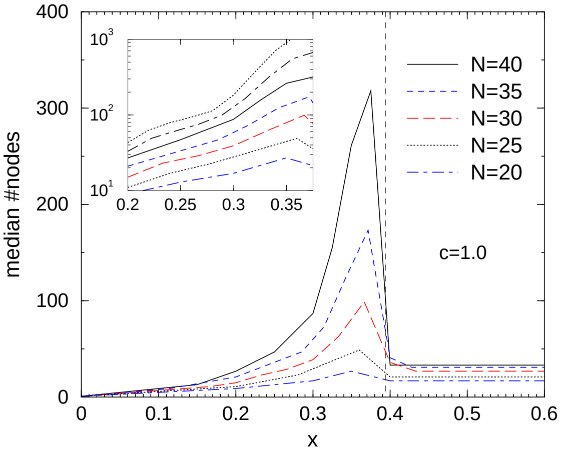

Figure 2: Time complexity of the vertex cover: Median number of nodes visited in the backtracking tree as a function of the fraction x of covered vertices for graph sizes $N = 20, 25, 30, 35, 40$ ($c = 1.0$). The inset shows the region below the threshold with logarithmic scale, including also data for $N = 45, 50$. The fact that in this representation the lines are equidistant demonstrates that the time complexity grows exponentially with N.

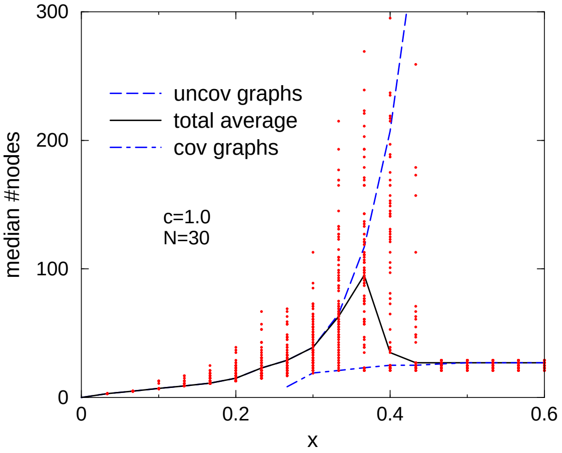

Figure 3: Median number of nodes visited in the backtracking tree as a function of the fraction x of covered vertices, displayed separately for the cases of coverable and uncoverable graphs ($N = 30, c = 1.0$). Additionally, a scatter plot of the number of nodes for 100 realizations is presented: for each run a dot is included in the figure.

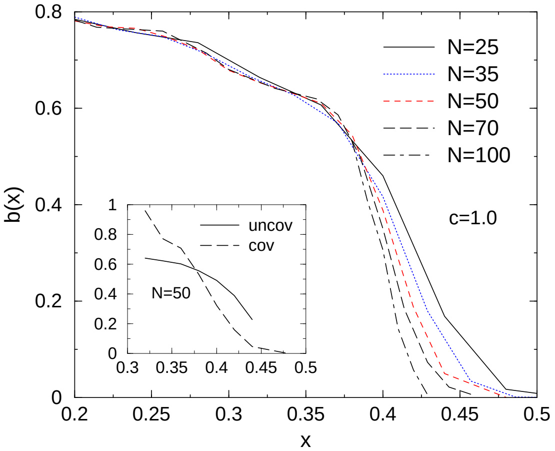

Figure 4: The fractional size $b(x)$ of the backbone as a function of the relative cardinality x of the vertex cover. The results are for the case $c = 1.0$ and for the system sizes $N = 25$, 35, 50, 70, and 100. The inset shows results for $N = 50$. There the fractional backbone sizes are displayed either for the subset of graphs which are coverable with xN vertices (cov) or for uncoverable graphs (uncov). The total function $b(x)$ is almost the minimum of both curves.

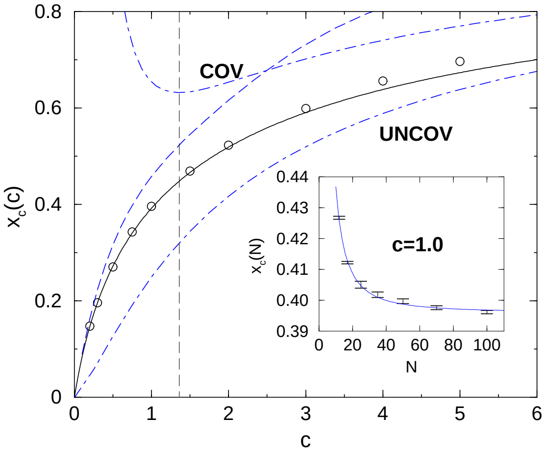

Figure 5: Phase diagram: critical fraction x_c of covered vertices as a function of the edge density c. For $x > x_c$, almost all graphs have covers with xN vertices, while they have almost surely no cover for $x < x_c$. The solid line shows the analytic result. The circles represent the results of the numerical simulations. Error bars are much smaller than symbol sizes. The upper bound of Harant is given by the dashed line, the bounds of Gazmuri by the dash-dotted lines. The vertical line is at $c = e/2$. Inset: All numerical values were calculated from finite-size scaling fits of $x_c(N, c)$ using functions $x_c(N) = x_c + aN^{-b}$. We show the data for $c = 1.0$ as an example.

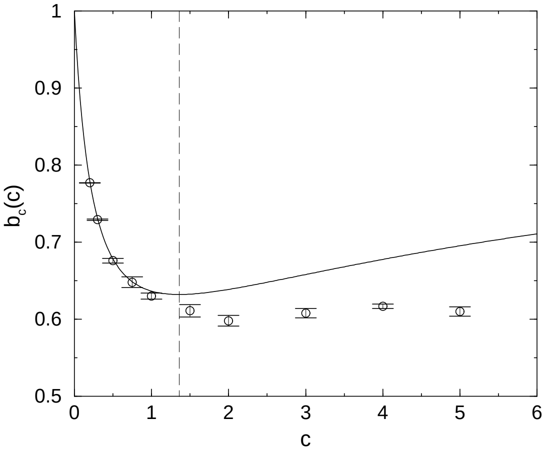

Figure 6: The backbone size b_c at the critical point as a function of c. The solid line shows the analytic result. The numerical results are represented by the error bars. They were obtained from finite-size scaling fits similar to the calculation for $x_c(c)$. The vertical line is at $c = e/2$.

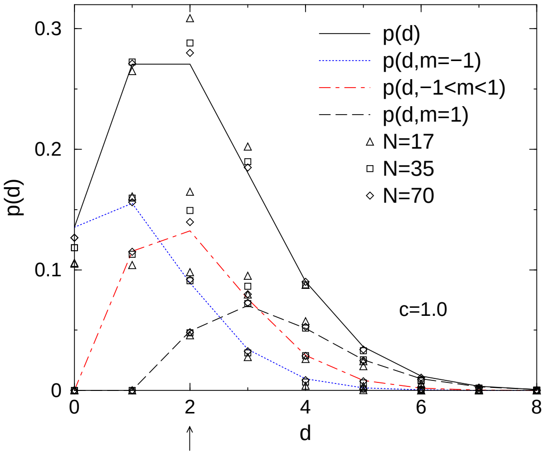

Figure 7: Distribution of degrees d at the threshold ($c = 1.0$). We show the total distribution of the degrees, determined by the ensemble of random graphs, as well as results describing the minimal vertex covers. The total distribution is divided into three contributions arising from the vertices which either are not in the backbone (magnetization $-1 < m < 1$) or which are in the backbone and have magnetizations $m = 1$ or $m = -1$. Analytical predictions are represented by the lines (which are guides to the eyes only, connecting the results for integer arguments), while the numerical results for $N = 17, 35, 70$ are displayed using the symbols.

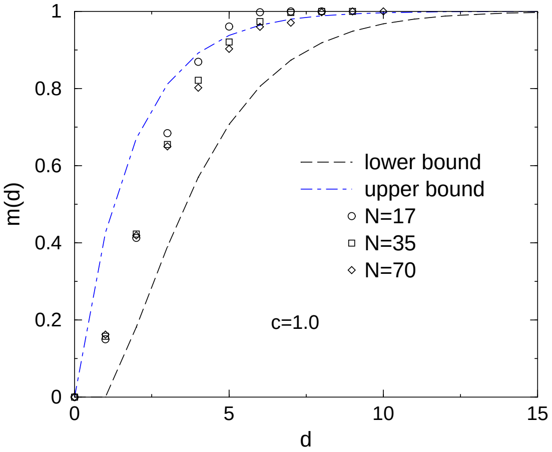

Figure 8: The average magnetization of a vertex at the threshold as a function of its degree d. The lower and upper bounds obtained from the analytical calculation in the $N \to \infty$ limes are shown by the lines. The symbols display the numerical results for $N = 17, 35, 70$.

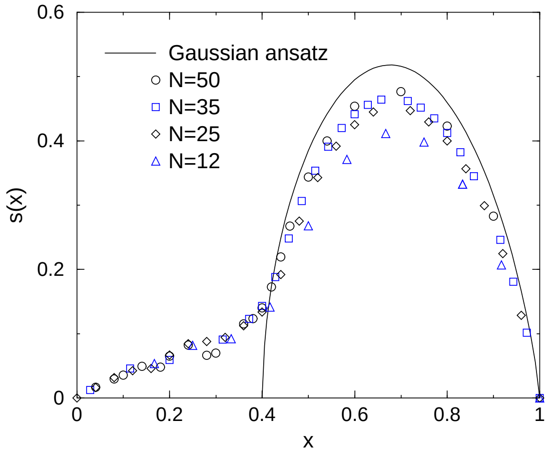

Figure 9: Entropy of the configurations as a function of the relative cardinality x of the vertex cover. The symbols represent results from the numerical enumerations, for different graph sizes N. The solid line displays the result from the Gaussian variational approximation.