ELSEVIER Artificial Intelligence 81 (1996) 155-181

Artificial Intelligence

Locating the phase transition in binary constraint satisfaction problems

Barbara M. Smith[a,*], Martin E. Dyer[b,1]

[a] Division of Artificial Intelligence, School of Computer Studies, University of Leeds, Leeds LS2 9JT, UK

[b] Division of Computer Science, School of Computer Studies, University of Leeds, Leeds LS2 9JT, UK

Received June 1994; revised February 1995

Abstract

The phase transition in binary constraint satisfaction problems, i.e. the transition from a region in which almost all problems have many solutions to a region in which almost all problems have no solutions, as the constraints become tighter, is investigated by examining the behaviour of samples of randomly-generated problems. In contrast to theoretical work, which is concerned with the asymptotic behaviour of problems as the number of variables becomes larger, this paper is concerned with the location of the phase transition in finite problems. The accuracy of a prediction based on the expected number of solutions is discussed; it is shown that the variance of the number of solutions can be used to set bounds on the phase transition and to indicate the accuracy of the prediction. A class of sparse problems, for which the prediction is known to be inaccurate, is considered in detail; it is shown that, for these problems, the phase transition depends on the topology of the constraint graph as well as on the tightness of the constraints.

Keywords: Search phase transitions; Constraint satisfaction; Crossover point; Mushy region; Expectation and variance of number of solutions

1. Introduction

Cheeseman, Kanefsky and Taylor [2] note that for many NP-complete or NP-hard problems, a phase transition can be seen as a control parameter is varied; the transition is from problems that are under-constrained, and so relatively easy to solve, to problems that are over-constrained, and so relatively easy to prove insoluble. They observed that

* Corresponding author. E-mail: bms@scs.leeds.ac.uk.

[1] E-mail: dyer@scs.leeds.ac.uk.

SSDI 0004-3702(95)00052-6

the problems which are on average hardest to solve occur between these two types of relatively easy problem, and further that, in the cases they considered, the phase transition becomes increasingly abrupt as problems become larger. For instance, for Hamiltonian Circuit problems, the order parameter giving the phase transition is the connectivity of the graph and the sharpness of the phase transition increases with graph size.

Williams and Hogg [13-15] have developed approximations to the cost of finding the first solution and to the probability that a problem is soluble, both for specific classes of constraint satisfaction problem (graph colouring, k-SAT) and for the general case. They show that in the limit as the number of variables becomes large, their approximations exhibit both a step change in the probability that a problem is soluble and a peak in the cost of finding the first solution, at the same critical value of the control parameter. Although Williams and Hogg show that their predictions of the critical value match the experimental data given in [2,7] reasonably well, their work is essentially based on the asymptotic behaviour of approximations, showing an instantaneous phase transition.

In finite problems, the phase transition is not instantaneous, but occurs over a range of values of the control parameter. This paper is concerned with the phase transition in finite constraint satisfaction problems (CSPs), the intention being to investigate not only the point at which the average cost of solving problems, or proving them insoluble, is greatest, but also the boundaries of the phase transition. Kirkpatrick and Selman [6] also consider phase transitions in finite problems, discussing the dependence of the width of the phase transition on problem size for k-SAT problems.

In phase transitions of the kind modelled by applied mathematicians, for instance from a solid to a liquid phase, the cause of the phase transition may be modelled by an instantaneous change in some environmental parameter. However, the *effect* of the change may take place over a finite spatial region; this region (in which the material is neither completely liquid nor completely solid) is referred to as the *mushy region*. The term is used in this paper to denote the range of values of the control parameter over which the phase transition from solubility to insolubility takes place, in order to emphasise that the transition is not instantaneous. The mushy region can be defined as the range of values of the control parameter over which the probability that a problem is soluble falls from 0.99 to 0.01 (choosing these limits arbitrarily) and approximated by the range of values over which the observed proportion of soluble problems falls from $\geqslant 99\%$ to $\leqslant 1\%$.

Mitchell, Selman and Levesque [7] carried out experiments with satisfiability problems in which they noted that, as the number of formulas in random clauses is varied, the hardest problems occur where 50% of the problems are satisfiable. Crawford and Auton [3] took this work further in order to predict the location of the 50% satisfiable point, which they term the *crossover point*. It will be assumed in this paper that, for CSPs in general, the crossover point and the maximum average solution cost coincide: there is a great deal of experimental evidence to support this assumption.

The paper begins by discussing the generation model used for the experimental CSPs. Section 3 describes the phase transition in a class of small CSPs; the behaviour of these CSPs suggest that the crossover point can be predicted using the expected number of solutions, and this is discussed in Section 4. Section 5 investigates in detail a class of sparse constraint graphs, for which the prediction is inaccurate: it is shown that in some cases, a precise prediction of the crossover point would need to take into account the characteristics of the constraint graph, in particular, the degree distribution. It is shown that there is a close linear correlation between the crossover point and the regularity of the graph; the correlation is even better if end-vertices in the graph are ignored. The main reason for the failure of a predictor based on the expected number of solutions is the very high variance. Section 6 discusses the variance of the number of solutions in more detail, and shows that the variance can indicate when the predictor can be expected to give good results. Finally the paper discusses applications to real CSPs.

2. Random binary CSPs

A (finite) constraint satisfaction problem consists of a finite set of variables $X = \{x_1, \ldots, x_n\}$; for each variable, a set D_i of possible values (its domain); and a set of constraints, each of which consists of a subset $\{x_i, \ldots, x_j\}$ of X and a relation $R \subseteq D_i \times \cdots \times D_j$; informally, the constraint specifies the allowed tuples of values for the variables it constrains (see for instance Tsang [12] for a detailed account of CSPs). A solution to a CSP is an assignment of a value from its domain to every variable such that all the constraints are satisfied; a constraint is satisfied if the tuple of values assigned to the variables it constrains is in the constraint relation. A k-ary constraint constrains k of the problem variables; in a binary CSP, all the constraints are binary.[2]

The constraint graph of a binary CSP is a graph in which there is a vertex representing each variable and for every constraint there is an edge linking the affected variables. A binary constraint relation between a pair of variables with m_1 and m_2 values in their respective domains can be represented by an $m_1 \times m_2$ matrix of boolean values; a "true" value indicates that the corresponding pair of values is allowed by the constraint.

For the experiments described below, sets of randomly-generated binary CSPs were used. Each set of problems is characterised by four parameters: n, the number of variables; m, the number of values in each variable's domain; p_1, the probability that there is a constraint between a pair of variables, and p_2, the conditional probability that a pair of values is inconsistent for a pair of variables, given that there is a constraint between the variables. The parameters p_1 and p_2 are the *constraint density* and the *constraint tightness*, respectively. Other work with randomly-generated CSPs (see for instance [4]) has also defined sets of problems in terms of quantities corresponding to these four parameters, albeit using different terminology.

There are several possible ways of treating the probabilities p_1 and p_2. One possibility is to select each of the $n(n-1)/2$ possible edges in the constraint graph independently with probability p_1, and then, for each pair of variables linked by a constraint, generate the relation matrix by assigning the value "false" to each pair of values independently with probability p_2. As far as generating the constraint graph is concerned, this corresponds to the model termed Model A by Palmer [8]. This method would give a

[2] Or unary, but since unary constraints can be dealt with by reducing the domains of the affected variables, these will be ignored in this paper.

set of problems in which, on average, the number of constrained pairs of variables is $p_1n(n-1)/2$, and for each pair of constrained variables, the average number of inconsistent pairs of values is m^2p_2. However, within a set of randomly-generated problems, there may be considerable variation.

Since the ultimate aim in considering sets of randomly-generated CSPs is to be able to make predictions about the behaviour of problems of a given size, with particular observed numbers of constraints and numbers of inconsistent pairs of values, it was decided to generate sets of problems in which p_1 and p_2 specify precisely, rather than on average, how many constraints and pairs of inconsistent values there should be. Hence, for each set of randomly-generated problems, there should be exactly $p_1n(n-1)/2$ constraints (rounded to the nearest integer), and for each pair of constrained variables, the number of inconsistent pairs of values should be exactly m^2p_2. To allocate the constraints, $p_1n(n-1)/2$ of the possible variable pairs are chosen at random; for each constrained pair of variables, m^2p_2 of the m^2 possible pairs of values are chosen at random. This model is used for all the experiments described in this paper. It is an extension of the random graph model referred to by Palmer as Model B, and will be referred to as Model B below.

In each of the series of experiments described below, n, m and p_1 were fixed, and p_2 was the varying control parameter; a series of experiments will be referred to by the tuple $\langle n, m, p_1\rangle$, and a set of random problems with the same four parameters will be referred to by $\langle n, m, p_1, p_2\rangle$. One minor disadvantage of Model B, compared with the first model, is that one cannot vary p_2 in steps of less than $1/m^2$; in the experiments, the value of m is 10, allowing p_2 to be varied in steps of 0.01.

3. A well-behaved case

A series of experiments was carried out with sets of randomly-generated binary CSPs, generated according to Model B, with $n = 8$, $m = 10$, $p_1 = 1.0$ and p_2 varying. These are small problems, and therefore do not exhibit as sharp a phase transition as has been observed in larger problems, for instance in [10]. However, one of the planned experiments was to find all solutions to the problems over a wide range of values of p_2 and this could not be done for large problems. As will be seen later, the behaviour of these problems is relatively well-behaved and so serves as a good starting point for investigating the phase transition in binary CSPs.

For these experiments, the randomly-generated CSPs were solved using the forward checking algorithm, using the fail-first principle to select the variable with smallest remaining domain as the next variable to be instantiated [5]. This algorithm is known to be reasonably efficient and can be used to find either just one solution, or all solutions. It was intended to be representative of its class of CSP algorithms, i.e. depth-first search algorithms which seek to extend consistent partial solutions, backtracking when failure is detected.

Fig. 1 shows the median cost, measured by the number of consistency checks required to find one solution or prove that there is no solution for each of a set of $\langle 8, 10, 1.0, p_2\rangle$ problems. The minimum and maximum cost observed in each set of problems is also

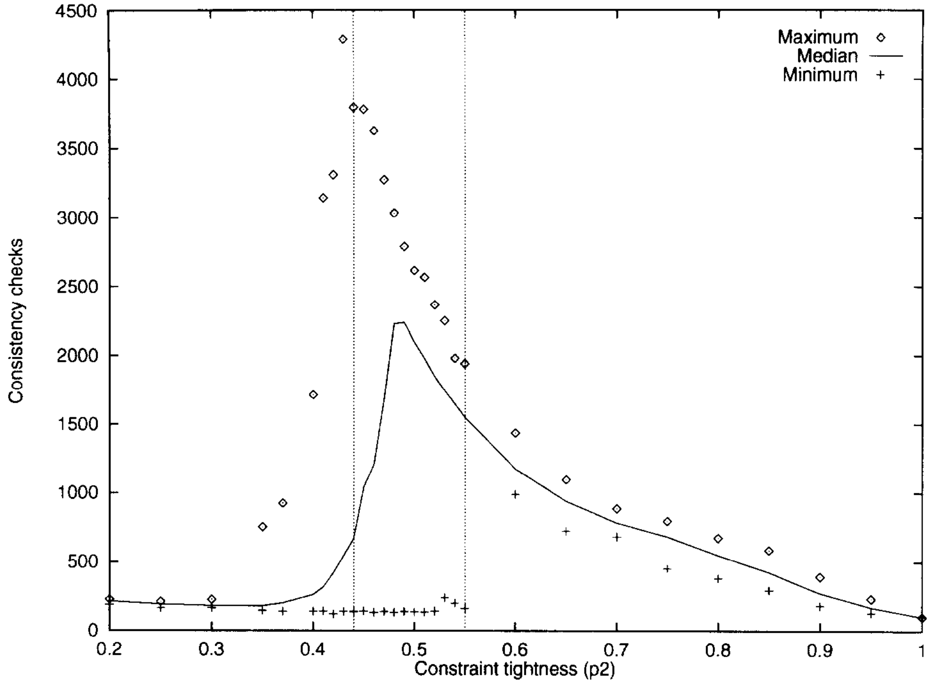

Fig. 1. Median cost (to find one solution or show that there are none) for CSPs with $n = 8$, $m = 10$, $p_1 = 1.0$.

shown. The vertical lines in Fig. 1 show the boundaries of the mushy region: the largest value of p_2 at which at least 99% of problems are soluble, and the smallest value at which not more than 1% are soluble. (For these particular samples, no insoluble problem occurs to the left of the mushy region and no soluble problem to the right.) In order to get a clear picture of the behaviour over the phase transition, 500 problems were generated for each value of p_2 between 0.44 and 0.55; smaller samples were required elsewhere, where the behaviour is much less variable.[3]

A notable feature of Fig. 1 is the peak in the median cost which occurs during the phase transition. (For these problems, the peak in the mean search effort occurs at the same value of p_2 as the peak in the median cost. However, for some less well-behaved sets of problems, individual problems with very high solution cost can occur at values of p_2 below the phase transition and can distort the mean cost. Hence the median has been used as the measure of the average cost throughout.)

The minimum cost at each value of p_2 remains very low as long as there are problems in the sample which have a solution; if a problem has at least one solution, there is always a chance that it will be found very quickly, and in these experiments the sample size at each value of p_2 was sufficiently large that this did in fact happen for at least one problem throughout the mushy region. Most of the maximum values in the mushy

[3] Solving 500 problems with these parameter values takes 52 CPU seconds when $p_2 = 0.48$ (using a C program running on a SPARCstation IPX).

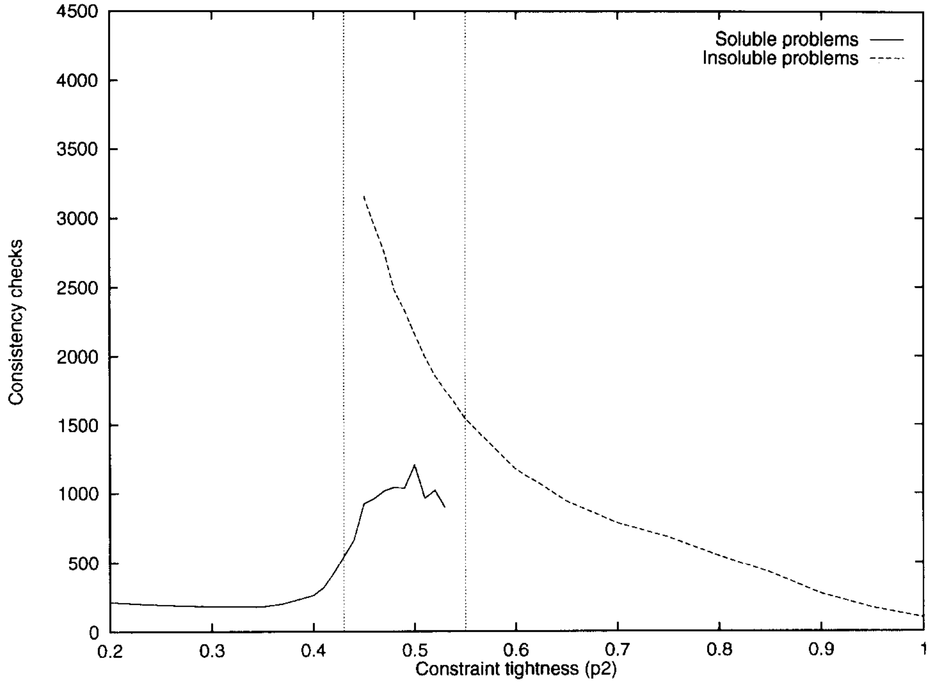

Fig. 2. Median cost for CSPs with $n = 8$, $m = 10$, $p_1 = 1.0$; soluble and insoluble problems shown separately.

region (and certainly those beyond it) are due to insoluble problems; however, the overall maximum value (at $p_2 = 0.43$) is clearly caused by a problem which does have solutions, since at that point all the problems have solutions. Fig. 1 illustrates what is commonly observed: that the solution cost is very variable around the peak in average cost, and that the greatest variability occurs before the peak.

The peak corresponds approximately to the value of p_2 at which half the problems are insoluble and half are soluble, the point referred to by Crawford and Auton [3] as the crossover point. In fact, exactly 50% of the problems have a solution at $p_2 = 0.48$; the median cost is almost identical at $p_2 = 0.48$ and 0.49. For smaller values of p_2, for which all problems are soluble, problems are much easier to solve, on average, until there is a sharp increase in difficulty as p_2 increases and insoluble problems begin to occur. For larger values of p_2, as problems become uniformly insoluble, the fall in the median consistency checks is much more gradual, and it is an over-simplification to describe this side of the phase transition as a region of easy problems. Although it does become easy to prove insolubility for p_2 close to 1, many of the problems in this region are easy only by comparison with the insoluble problems occurring in the mushy region.

Fig. 2 shows the same set of problems as Fig. 1, but with soluble problems separated from insoluble problems in the mushy region, where a mixture of soluble and insoluble problems occurs. (A similar graph displaying satisfiable and unsatisfiable 3-SAT problems separately is given in [7].)

At the edges of the phase transition, the population of problems is dominated by one of these two types of problem, so that some of the points on the graph represent only small numbers of problems (although points representing fewer than 20 problems have been omitted). Separating soluble from insoluble problems in this way leads to a plausible explanation for the fact that the maximum average search cost occurs during the phase transition, for any algorithm of the same general type as forward checking. For insoluble problems, the search effort decreases as p_2 increases, because the increasingly tight constraints allow a greater degree of pruning in any algorithm which backtracks as soon as it encounters a failure. Hence the cost, for insoluble problems, is greatest at the smallest value of p_2 for which insoluble problems occur, i.e. in the mushy region.

The case of soluble problems is more complex. As p_2 increases, the fact that the number of solutions is decreasing rapidly becomes significant and it becomes harder (for any algorithm) to find a solution. During the phase transition, the soluble problems have very few solutions and as solutions become rarer, the algorithm must on average explore more of the induced search space before finding the first solution. Fig. 2 shows that the search effort in fact appears to reach a maximum and begins to decline, just as the soluble problems are running out; this can be explained by arguing that for problems with only one solution, which for this sample increasingly dominate as soluble problems become scarce, the algorithm must on average explore half the search space before finding the solution, and since the total size of the search space decreases as p_2 increases, the search effort to find a single solution similarly starts to decrease, just before the soluble problems disappear altogether. Fig. 2 is consistent with the experiments on graph colouring problems reported by Cheeseman, Kanefsky and Taylor [2], where a peak in the average solution cost of soluble problems was found, as the average connectivity increased. In that case, the problems were generated in such a way that they were guaranteed to have a solution, and thus very easy soluble problems were found well beyond the peak in average cost, in contrast to Fig. 2. This suggests that the decrease in average solution cost beyond the crossover point shown in Fig. 2 would continue as p_2 increases.

To summarise, soluble problems are easier to solve as p_2 decreases from the crossover point; insoluble problems are easier to prove insoluble as p_2 increases, and overall the maximum average search effort must occur in the mushy region, where the most difficult soluble problems and the most difficult insoluble problems co-exist. This is exemplified in Figs. 1 and 2.

As problems get very large ($n \to \infty$), experimental evidence suggests that the phase transition becomes increasingly sharp, so that in the limit there is an instantaneous change from soluble to insoluble problems at a single value of p_2, and we should expect that the maximal search effort will then coincide with this point. However, for finite problems, the phase transition occurs over a range of values of p_2 (defining the mushy region), and $p_{2\text{crit}}$, the value at which the average search effort reaches a maximum and (by assumption) the probability that a problem has a solution is 0.5, occurs at some point in that range.

Clearly the phase transition is independent of the algorithm used to find the solutions, though the number of consistency checks required to find the first solution, or to show that there is no solution, depends very much on the algorithm. However, Prosser's

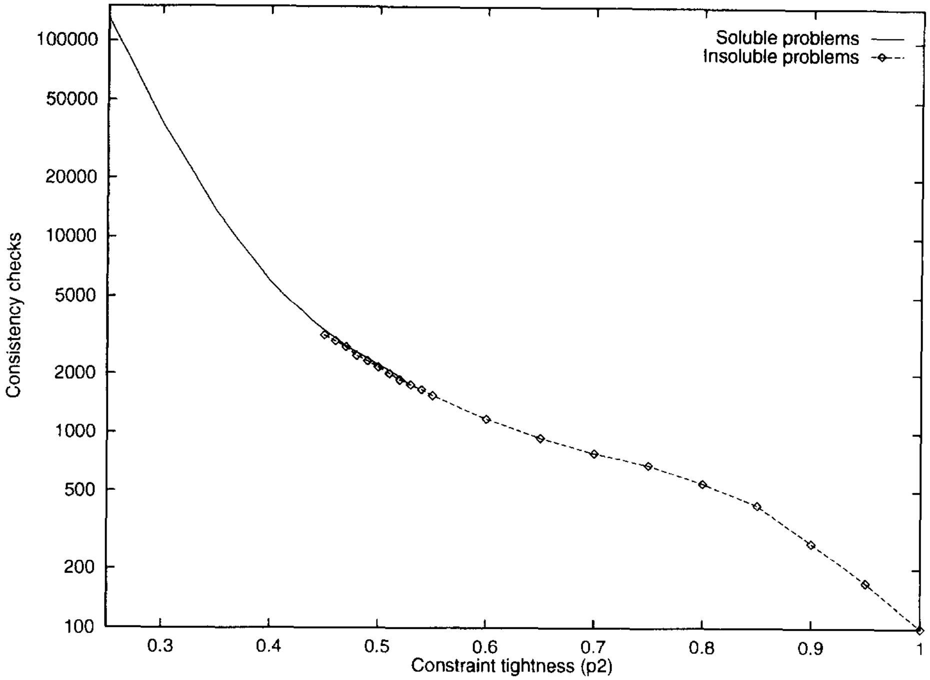

Fig. 3. Median cost (to find all solutions or show that there are none) for CSPs with $n = 8$, $m = 10$, $p_1 = 1.0$.

experimental results [10] for a CSP algorithm which combines forward checking with conflict-directed backjumping, described in [9], show similar qualitative behaviour to that shown in Fig. 1. In particular the peak in the median consistency checks appears to occur at the same value of p_2 for different algorithms. The results reported in [10] also show that the qualitative behaviour of larger problems (for instance $n = 20$, $m = 20$; $n = 30$, $m = 10$), and for a range of values of p_1, is similar to that shown in Fig. 1, although the peak becomes much less sharply defined for small values of p_1. It appears therefore that the phase transition behaviour shown in Fig. 1 is common to constraint satisfaction problems in general when subjected to backtracking depth-first search algorithms. Similar curves are also shown in [7] for randomly-generated 3-SAT problems.

As well as the first solution, all solutions were found for the sample problems shown in Fig. 1. It is instructive to plot the median cost to find all solutions for the soluble problems, and compare the results with the median cost to prove insolubility for the insoluble problems. The two curves are shown in Fig. 3; by definition, they overlap in the mushy region, where there are both soluble and insoluble problems. (The curve representing insoluble problems is identical to the right hand curve of Fig. 2, but with a log scale on the vertical axis.)

It can be seen from Fig. 3 that the two curves are virtually indistinguishable in the mushy region; for a given value of p_2 it appears to be neither more nor less difficult to find all solutions to the soluble problems than to show that the insoluble problems have no solutions. (In fact, it requires fewer consistency checks on average to prove insolubility than to find all solutions at the same value of p_2, but the difference is very small for these problems and hence is barely detectable in Fig. 3.)

If we require to find all solutions to a CSP, or prove that there are none, the median cost decreases smoothly and rapidly as p_2 increases, and nothing noteworthy happens as the problems become insoluble. The phase transition is only an interesting event if just one solution is required: it can then also be viewed as a transition from a partial search of the induced search space (which can be terminated as soon as a solution is found) to a complete search (which is required if there are no solutions). The transition therefore involves a more or less sudden jump from the lefthand curve of Fig. 2 to the righthand curve.

4. The expected number of solutions

In the previous section, the phase transition and its associated phenomena were described, at least for one particular set of parameters. It would be useful to know over what range of values of p_2 the mushy region occurs, and where the crossover point is, without extensive experimentation. However, we have no way, so far, of accurately estimating the probability that a binary CSP is soluble.

For the problems discussed in the last section, at the crossover point, i.e. where 50% of problems have a solution, the soluble problems have very few solutions. Of the 500 random problems generated at $p_2 = 0.48$, the 250 soluble problems have, on average, 2 solutions (and therefore the average number of solutions for all 500 problems is 1). If this observation, that at the crossover point the soluble problems have very few solutions, is generally true, then instead of trying to estimate the probability that a problem is soluble directly, we can alternatively look for a value of p_2 at which the expected number of solutions is small.

It is possible to do this, since, for CSPs generated following Model B, the expected number of solutions, $E(N)$, is given by:

$$E(N) = m^n(1 - p_2)^{n(n-1)p_1/2}, \tag{1}$$

i.e. the number of possible assignments of m values to n variables, multiplied by the probability that a randomly-chosen assignment is consistent. (See the appendix for a derivation of Eq. (1)). For instance, when $p_2 = 0$ (or $p_1 = 0$), there are no inconsistent pairs of values and $E(N) = m^n$; when $p_2 = 1$ (and $p_1 > 0$) there are no solutions. The expected number of solutions decreases very rapidly, from m^n, as p_2 increases from zero. The sample problems discussed in Section 3, for which all solutions were found, as well as other experimental results, confirm that (1) gives very accurate results for the average number of solutions, given a large sample of problems. The formula can easily be modified to allow p_2 to take different values for different constraints and to allow different domain sizes for different variables, if this is a more realistic model for a particular CSP.

For a well-behaved case, such as that discussed in the previous section, a value of p_2 for which $E(N)$ is small, but not too close to zero, can be expected to give a mixture of problems with no solutions and problems with very few solutions, i.e. a point in the mushy region, and therefore close to the crossover point, $p_{2\mathrm{crit}}$. Experimental results suggest that, as for the $\langle 8, 10, 1\rangle$ problems, $E(N) = 1$ would give a good predictor, $\hat{p}_{2\mathrm{crit}}$, of the crossover point.

From $E(N) = 1$, we have $m^n(1 - \hat{p}_{2\mathrm{crit}})^{n(n-1)p_1/2} = 1$, and hence:

$$\hat{p}_{2\mathrm{crit}} = 1 - m^{-2/((n-1)p_1)}. \tag{2}$$

For the $\langle 8, 10, 1.0\rangle$ problems of the last section, the value of $E(N)$ when $p_2 = 0.48$ is 1.12, and $\hat{p}_{2\mathrm{crit}}$ is 0.482, corresponding slightly better to the observed peak in the median search effort, which appears to occur between 0.48 and 0.49.

Prosser [10] gives results comparing the observed values of $p_{2\mathrm{crit}}$ (in this case, the observed peak median cost) with the estimated values given by (2) for three series of experiments: $n = 20$, $m = 10$; $n = 20$, $m = 20$; and $n = 30$, $m = 10$. The experimental results show that the observed value of $p_{2\mathrm{crit}}$ and the predicted value $\hat{p}_{2\mathrm{crit}}$ are in close agreement, except for low values of p_1 (smaller than 0.3), when $\hat{p}_{2\mathrm{crit}}$ is an over-estimate of $p_{2\mathrm{crit}}$. That is, for small values of p_1, $E(N)$ is greater than 1 at the crossover point. This discrepancy will be discussed further below.

The assertion that the point at which $E(N) = 1$ marks the phase transition, from a region where most problems have many solutions to a region where most problems have no solutions, is also made by Williams and Hogg [15]. They note that the choice of parameters which makes $E(N) = 1$ marks the boundary between a region in which $E(N)$ increases exponentially with n (the number of variables in the problem) and a region in which $E(N)$ decays exponentially with n.[4]

If the expected number of solutions is very small, then, as Williams and Hogg point out, we can safely conclude that the probability that a problem has any solutions, p_{sol}, is likewise very small, from the Markov inequality, which gives:

$$p_{\mathrm{sol}} = P(N \geqslant 1) \leqslant E(N). \tag{3}$$

For large n, $E(N) \to 0$ for all $p_2 > \hat{p}_{2\mathrm{crit}}$ and so $p_{\mathrm{sol}} \to 0$.

However, we cannot similarly assume that, if the expected number of solutions is very large, most problems have many solutions, or even that most problems have any solutions: it depends on the variance, $\mathrm{var}(N)$. A bound on p_{sol} is given by the Cauchy inequality [1]:

$$1 - p_{\mathrm{sol}} = P(N = 0) \leqslant \frac{\mathrm{var}(N)}{E(N)^2 + \mathrm{var}(N)}. \tag{4}$$

Making the assumption that $\mathrm{var}(N)/E(N)^2 \to 0$ as $n \to \infty$, when $E(N) > 1$, Williams and Hogg conclude that $p_{\mathrm{sol}} \to 1$ and hence that, asymptotically, there is an instantaneous phase transition at the point where $E(N) = 1$, i.e. at $\hat{p}_{2\mathrm{crit}}$.

Since for the $\langle 8, 10, 1\rangle$ experiments described in Section 3, the point at which $E(N) = 1$ does correspond very well to the crossover point, where $p_{\mathrm{sol}} = 0.5$, it is clear that at

[4] Eq. (1) can be written as $E(N) = [m(1 - p_2)^{(n-1)p_1/2}]^n$, and at $\hat{p}_{2\mathrm{crit}}$ the term inside the square brackets is equal to 1.

that point var(N) must be small. The behaviour of var(N) in other cases is discussed in Section 6.

5. Sparse constraint graphs

As already noted, Prosser [10] observed that $\hat{p}_{2\mathrm{crit}}$, the value of p_2 at which $E(N) = 1$, is a good predictor of the location of the peak median cost for most of the CSPs he studied, but not for those with the sparsest constraint graphs, with $p_1 < 0.3$. Furthermore, he found that in some cases $\hat{p}_{2\mathrm{crit}}$ is not even in the mushy region; there were no soluble problems in his samples, even though at that point $E(N) = 1$. It was decided to investigate the behaviour of these sparse problems in more detail: this section presents a description of the phase transition behaviour of a set of sparse CSPs, in order to explain why the arguments presented earlier break down in these cases.

It was at first thought that the discrepancy between $\hat{p}_{2\mathrm{crit}}$ and $p_{2\mathrm{crit}}$ might be partly caused by disconnected graphs, and in particular by graphs with isolated vertices. If the constraint graph has i isolated vertices (corresponding to unconstrained variables) and the rest of the problem has exactly one solution, then the whole problem has m^i solutions. Hence, problems which are "only just" soluble will have many solutions, rather than very few as in Section 3. To eliminate this factor, it was decided to consider only connected graphs. Problems were generated using Model B, as before, but each constraint graph was checked for connectedness: any disconnected graph was thrown away and a new graph was generated. The results presented in the rest of this section refer to connected graphs only; however, $\hat{p}_{2\mathrm{crit}}$ is still not a good predictor of the phase transition for small n when p_1 is also small.

Fig. 4 shows the phase transition for three sets of problems with $n = 30$, $m = 10$, and increasingly sparse constraint graphs. For each value of p_2, 100 problems (with connected constraint graphs) were solved. As before, the vertical lines mark the boundaries of the respective mushy regions. For these sets of problems, $\hat{p}_{2\mathrm{crit}}$ is 0.41 for $p_1 = 0.3$, 0.55 for $p_1 = 0.2$ and 0.79 for $p_1 = 0.1$. So $\hat{p}_{2\mathrm{crit}}$ is a worse predictor of the location of the peak median cost as p_1 gets smaller, and for $p_1 = 0.1$ is not even in the mushy region, even though the mushy region is getting wider as p_1 gets smaller. The peak median cost for the $\langle 30, 10, 0.1\rangle$ problems occurs for $p_2 = 0.73$, at which point 61% of the sample problems are soluble;[5] at this point, the expected number of solutions is 95,500.

In investigating the behaviour of these problems, an obvious difference from the well-behaved problems of Section 3 is that a random sample of $\langle 30, 10, 0.1\rangle$ problems contains a variety of constraint graphs, even when disconnected graphs are excluded, whereas the CSPs discussed in Section 3 all have the same constraint graph, i.e. the complete graph, K_8, since $p_1 = 1$. The question naturally arises whether different constraint graphs yield different behaviours.

[5] This is consistent with the peak occurring at the crossover point, since for $p_2 = 0.72$, 41% of the sample problems are soluble.

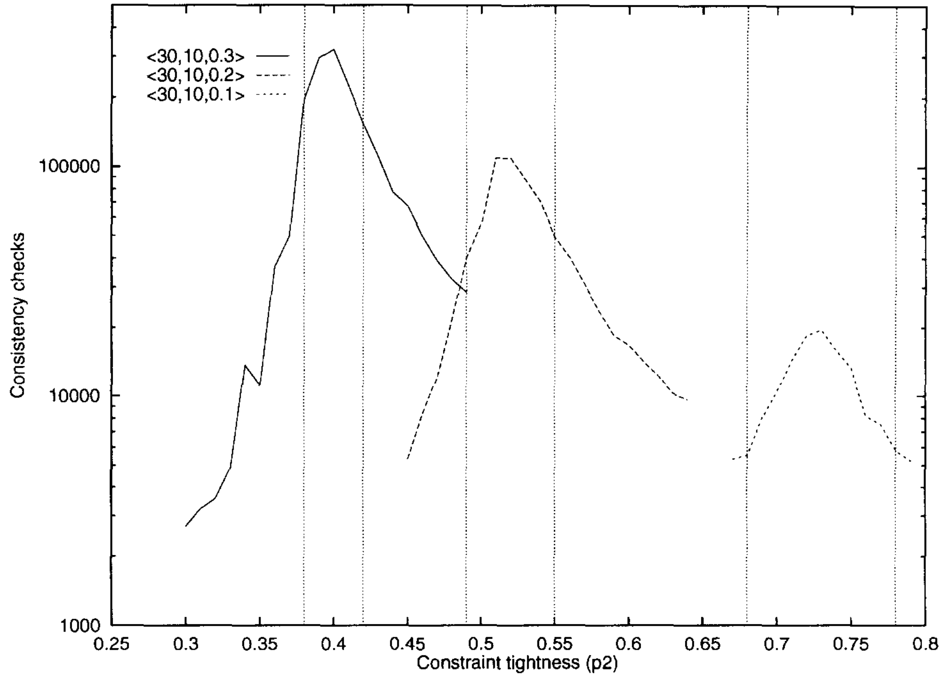

Fig. 4. Phase transition for three sets of sparse CSPs with $n = 30$.

The generation of a random instance of a CSP can be divided into two stages; first the generation of the constraint graph, governed by the parameter p_1, and secondly, the generation of the relation matrix for each pair of variables linked by a constraint, using p_2. It is possible to generate a single constraint graph and use it as the basis of a whole population of CSPs. In just the same way, the $\langle 8, 10, 1\rangle$ CSPs of Section 3 are problems with the same constraint graph, K_8. Hence, in carrying out experiments with CSPs whose constraint density p_1 is less than 1, there is a choice between using a separate randomly-generated constraint graph for each individual problem, or generating all of the required problems with the same constraint graph, which effectively becomes a fifth parameter, along with n, m, p_1 and p_2.

To investigate whether different constraint graphs give rise to different phase transition behaviours, a number of sets of CSPs were generated, with $n = 30$, $m = 10$, $p_1 = 0.1$ and several values of p_2 in the region of the phase transition, in such a way that all the problems within a set had the same constraint graph, but different sets had different constraint graphs. The proportion of soluble problems for a given value of p_2 was found to vary from set to set. Hence, the probability that a problem is soluble, p_{sol}, depends on the constraint graph, as well as on the four parameters n, m, p_1 and p_2.

This is perhaps intuitively obvious; it is at least easy to imagine that CSPs based on some constraint graphs would be less likely to have a solution than others with the same parameters. If a graph has one or more vertices of higher than average degree, i.e. with a large number of adjacent vertices, then the corresponding variable is highly

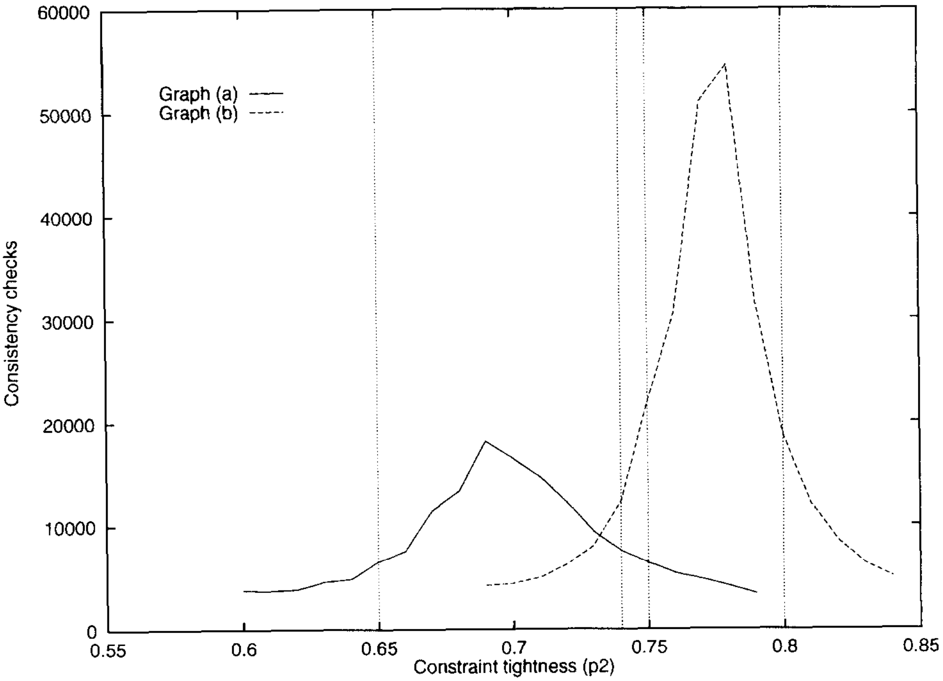

Fig. 5. Phase transition for two sets of $\langle 30, 10, 0.1\rangle$ CPSs with different constraint graphs.

constrained and would be difficult to assign a locally-consistent value. On the other hand, local consistency is more likely to be achievable in a graph with the same number of constraints, but much closer to being regular, i.e. with all vertices having approximately the same degree.

If p_{sol} at a particular value of p_2 depends on the constraint graph, then so does the crossover point; hence, each constraint graph has its own phase transition. This is demonstrated in Fig. 5, which shows the phase transition for two sets of $\langle 30, 10, 0.1\rangle$ problems with different graphs. Graph (a) is from a set of randomly-generated graphs and has a very irregular degree distribution. Graph (b) was hand-generated and is as close to regular as possible (28 vertices have degree 3, 2 have degree 2, giving the required total of 44 constraints). As expected, at any value of p_2, CSPs with the more regular graph are more likely to have a solution than those with the irregular graph, and so the crossover point for graph (b) occurs at a higher value of p_2. Fig. 5 also shows that the effort required to solve a problem or to show that it has no solution also depends on the constraint graph, at least during the phase transition.

The graphs shown in Fig. 5 are extreme cases, and do not show that there is a correlation between the regularity of the constraint graph and the location of the crossover point, in general. To establish whether or not this is the case, the relationship was investigated further for the $\langle 30, 10, 0.1\rangle$ problems.

Using Model B with $m = 10$, p_2 can only be varied in steps of 0.01, and it was decided to estimate the crossover point more precisely by linear interpolation from the proportion of soluble problems found at two adjacent values of p_2. For instance, if 35% of problems are soluble at $p_2 = 0.74$ and 65% are soluble at $p_2 = 0.75$, the crossover point is estimated to be 0.745. (For the experiments discussed in this section, it is almost invariably the case that the peak median cost occurs at a value of p_2 adjacent to the crossover point estimated in this fashion.)

In a population of constraint graphs with fixed n and p_1, the number of constraints, c, and so the number of edges, is fixed. The mean vertex degree is also constant $(2c/n)$. A very irregular graph has some vertices with much higher than average degree, and correspondingly other vertices with lower than average degree. The regularity might therefore be measured by the variance of the degree distribution, or, since the mean degree is constant, by $\sum_{i=1}^{n} d_i^2$, where d_i is the degree of vertex i.

The regularity of the constraint graph is indeed closely correlated with the crossover point. The correlation is slightly better if any end-vertices (vertices of degree 1) are first eliminated from the graph. It can be argued that an end-vertex has very little influence on whether or not the variable represented by the adjacent vertex can be found a consistent value. If a vertex of high degree is adjacent to several end-vertices, its corresponding variable is not as highly-constrained as the degree suggests. Hence, end-vertices should be ignored in calculating the contribution of their adjacent vertices to the regularity of the graph, if, as here, the regularity of the graph is being used as an indication of the likelihood that there is a solution to a CSP based on the graph. End-vertices are eliminated recursively, so that any vertex adjacent only to end-vertices and one other vertex are also eliminated.[6]

$\sum_{i=1}^{n'} d_i^2/n'$ was then calculated for the remaining graph, where n' is the number of remaining vertices. A large value of this measure indicates a very irregular graph in which some vertices are adjacent to many others, which in turn are adjacent to yet other vertices; the high-degree vertices correspond to variables which it may be difficult to find consistent values for. A small value of the regularity measure indicates that the graph is close to being regular, and all vertices have similar degree.

Fig. 6 shows the results: the crossover point is estimated from samples of 500 problems at each value of p_2.[7] The random graphs are a set of 30 randomly-generated graphs, intended to show the distribution of crossover points likely to occur, while the extreme graphs were picked from a much larger sample to show the extremes of regularity that might occur, together with graph (b) from Fig. 5, which is the point on the extreme left. Graph (a) from Fig. 5 is the point on the extreme right. There is clearly a close correlation between the regularity of the constraint graph, measured as described, and the crossover point for these problems. (The linear correlation coefficient is −0.951 for the data shown in Fig. 6. This is slightly better than if the end-vertices are not removed (correlation coefficient −0.926)).

[6] This procedure would cause any constraint graph which is a tree to disappear completely. On the other hand, CSPs whose constraint graphs are trees are known to be easy [4], so this may be sensible.

[7] For these parameter values, solving 500 problems takes approximately 175-350 CPU seconds, depending on the constraint graph (using a C program running on a SPARCstation IPX), and estimating the crossover point needs runs with at least two values of p_2, so that Fig. 6 required more than 20,000 CPU seconds in total.

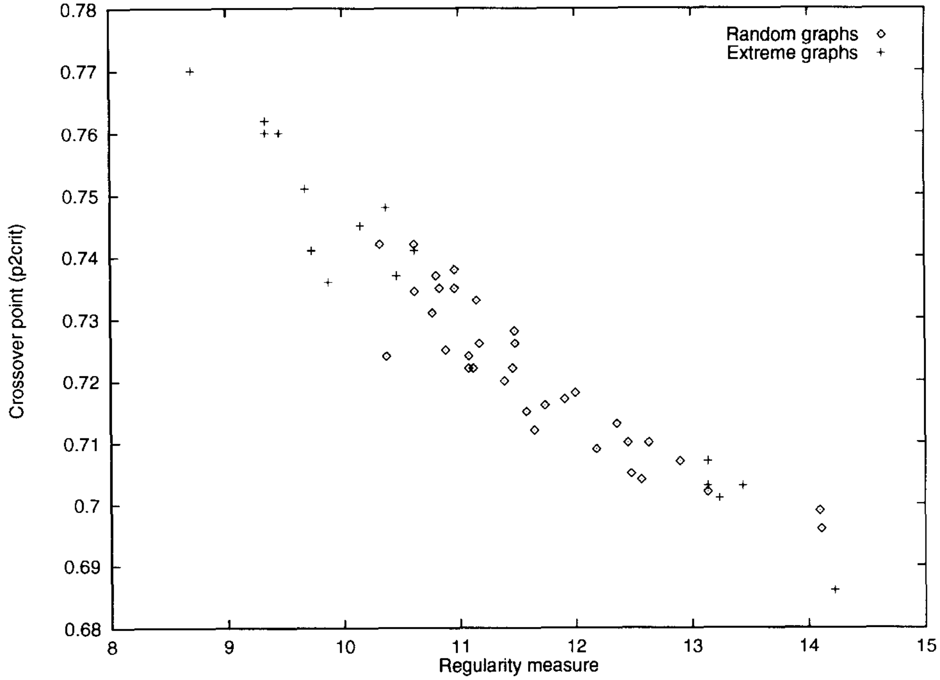

Fig. 6. Comparison between crossover point and regularity constraint graph for a set of $\langle 30, 10, 0.1\rangle$ CSPs.

It should be pointed out that these high correlation coefficients are very dependent on the way in which the regularity of the constraint graph is calculated. Nevertheless, the correlation is striking, and the removal of end-vertices makes a noticeable difference. It is possible that by identifying other relevant features of the constraint graph, a still better correlation could be achieved.

The random graphs from Fig. 6 suggest an explanation for the width of the mushy region for $\langle 30, 10, 0.1\rangle$, as shown in Fig. 4. The mushy region is based on a composite of many different constraint graphs whose crossover points occur, typically, anywhere between 0.7 and 0.74.

This evidence suggests that it might be difficult to predict the crossover point for small problems with sparse constraint graphs solely in terms of the parameters n, m and p_1, since the topology of the constraint graph also needs to be taken into account. However, experimental evidence ([11] and Fig. 4) suggests that, for larger values of n and/or p_1, the mushy region is narrower, and so the effect of the different constraint graphs will be less important.

Consideration of the effect of different constraint graphs does not explain, however, why the predicted crossover point at 0.79 is so inaccurate for the $\langle 30, 10, 0.1\rangle$ problems; it over-estimates the crossover point even for the regular graph, which from Fig. 6 probably has the highest possible crossover point for these problems. The discrepancy is due to the very high variance in the number of solutions for these problems: at $p_2 = 0.79$, the expected number of solutions is 1, but in practice almost all problems are insoluble. 1000 problems (with a different (connected) constraint graph for each one) were solved and yielded just one soluble problem, with 2340 solutions: hence the expected number of solutions is made up of a very high proportion of insoluble problems and a very small proportion of problems with many solutions. Conversely, at the observed crossover point (0.73), where $E(N) = 95{,}500$, the problems which are soluble (approximately half of the total) must have approximately 191,000 solutions on average (although this has not been verified experimentally).

Although p_{sol} depends on the constraint graph, $E(N)$ does not. This has been empirically verified by finding all solutions to samples of $\langle 30, 10, 0.1\rangle$ problems with two different connected constraint graphs at $p_2 = 0.77$. The constraint graphs were the regular graph (graph (b) of Fig. 5), for which 0.77 is the crossover point, and one of the random graphs from Fig. 6, whose crossover point was at 0.728. 200 problems based on the regular graph were solved; 51.5% have no solutions and the average number of solutions is 83.3, comparing well with $E(N)$ which is 82.4. For the other constraint graph, hardly any problems have solutions, as expected so far from its crossover point; 10,000 problems had to be solved in order to find a reasonable number of soluble problems. 99.1% of the problems have no solutions and the average number of solutions is 69.3; the 88 soluble problems have on average 7875 solutions.

Whether or not the constraint graph is taken into account, the expected number of solutions at the observed crossover point is much greater than the predictor $\hat{p}_{2\text{crit}}$ requires. Since, by definition, 50% of problems have no solution at the crossover point, this implies that the variance of the number of solutions is extremely high. The variance must also be very high at the predicted crossover point, since the rare problems that have solutions at that point have very many of them, to give an overall average of 1 solution. The variance of the number of solutions for this and other classes of CSP is discussed in more detail in the next section.

6. The variance of the number of solutions

In the previous section it was shown that for $\langle 30, 10, 0.1\rangle$ problems, the predictor of the crossover point given by (2) is very inaccurate, since $E(N) = 1$ does not indicate equal proportions of insoluble and soluble problems, as it does for the $\langle 8, 10, 1\rangle$ problems of Section 3, but that almost all problems are insoluble and the very few soluble problems have many solutions. It should be expected, therefore, that the variance of the number of solutions, $\text{var}(N)$, at $\hat{p}_{2\text{crit}}$ should be small for the $\langle 8, 10, 1\rangle$ problems and very large for the $\langle 30, 10, 0.1\rangle$ problems. If $\text{var}(N)$ can be calculated, then its value at $\hat{p}_{2\text{crit}}$ allows an informal check on the likely accuracy of the prediction of the crossover point. Furthermore, the inequalities (3) and (4) give bounds on p_{sol} at any value of p_2, provided that $E(N)$ and $\text{var}(N)$ are known.

For problems generated according to Model B, the variance can be calculated from:

$$E(N^2) =$$

$$m^n \sum_{s=0}^{n} \binom{n}{s} (m-1)^{n-s} \sum_{t=0}^{c} \frac{\binom{\binom{s}{2}}{t}\binom{\binom{n}{2}-\binom{s}{2}}{c-t}}{\binom{\binom{n}{2}}{c}} (1-p_2)^c \left(1 - \frac{m^2 p_2}{m^2-1}\right)^{c-t}, \quad (5)$$

where c is the number of constraints ($\binom{n}{2}p_1$, rounded to the nearest integer). The derivation of Eq. (5) is given in the appendix.

The value of this can be computed, with some effort, for moderate values of n. As expected, var(N) is small at $\hat{p}_{2\text{crit}}$ for $\langle 8, 10, 1\rangle$ problems (var(N) = 2.55 at p_2 = 0.482), but extremely large for $\langle 30, 10, 0.1\rangle$ problems at both the predicted and observed crossover points (var(N) = 1.16×10^6 at $p_2 = 0.79$; var(N) = 2.44×10^{12} at $p_2 = 0.73$).

The bounds given by (3) and (4) can be used to give lower and upper bounds on the mushy region. From the definition of the mushy region given earlier, the lower boundary is at $1 - p_{\text{sol}} = 0.01$ and the upper boundary is at $p_{\text{sol}} = 0.01$. From (3) and (4), bounds on the mushy region are given by the largest value of p_2 for which

$$\frac{\text{var}(N)}{E(N)^2 + \text{var}(N)} \leqslant 0.01 \tag{6}$$

and the smallest value of p_2 for which

$$E(N) \leqslant 0.01. \tag{7}$$

These inequalities cannot be expected to give very tight bounds on the mushy region for small n, but they may converge as n increases. Bounds on the crossover point can also be found by substituting 0.5 for 0.01 in (6) and (7). It has not been possible to calculate the variance of the number of solutions for large n, but assuming that the trends seen for smaller values of n continue, some conclusions can be drawn.

Fig. 7 shows the calculated bounds on the mushy region and the crossover point for a case in which they converge as n gets larger: here n and m are equal, and $p_1 = 1$. As n increases, it rapidly becomes prohibitively time-consuming to solve large samples of problems with these parameters, and empirical results are not available for $n > 20$. Prosser [10] showed that $p_{2\text{crit}}$ for $\langle 20, 20, 1\rangle$ problems is 0.27 (identical to $\hat{p}_{2\text{crit}}$); the mushy region for these problems is already very narrow. Fig. 7 confirms that for this class of problems, $\hat{p}_{2\text{crit}}$ is an accurate predictor of the crossover point for $n \geqslant 10$, and for $n > 70$, say, marks an almost instantaneous phase transition.

The behaviour of var(N) at $\hat{p}_{2\text{crit}}$ has also been investigated: informally, a small value of var(N) should indicate that $\hat{p}_{2\text{crit}}$ will be a reliable predictor of the crossover point, as for the $\langle 8, 10, 1\rangle$ problems; a very large value will indicate that it is likely to be an over-estimate, as for the $\langle 30, 10, 0.1\rangle$ problems. Table 1 shows the results for problems with $p_1 = 1$.

As can be seen from Table 1, for $\langle n, n, 1\rangle$, the class already described and shown in Fig. 7, var(N) at $\hat{p}_{2\text{crit}}$ appears to be small for all n, and in fact to decrease as n gets larger. The appendix gives an asymptotic analysis of (5) in this case, which shows that as $n \to \infty$, var(N) $\to$ e, which is consistent with Table 1.

Unfortunately, it is difficult to analyse the asymptotic behaviour of (5) in other cases. However, Fig. 8 shows the behaviour of var(N) at $\hat{p}_{2\text{crit}}$ for $\langle n, n, p_1\rangle$ problems, over a range of values of p_1. It appears that for these problems, var(N) at $\hat{p}_{2\text{crit}}$ decreases as n increases, from some point onwards (for $p_1 = 1$, the initial increase (if any) occurs for $n < 10$, and so is not shown in Table 1). For all values of n, it appears that var(N) is small (< 11.4) at $\hat{p}_{2\text{crit}}$ for $p_1 \geqslant 0.5$: this indicates that $\hat{p}_{2\text{crit}}$ will be an accurate estimate of the crossover point for all these problems.

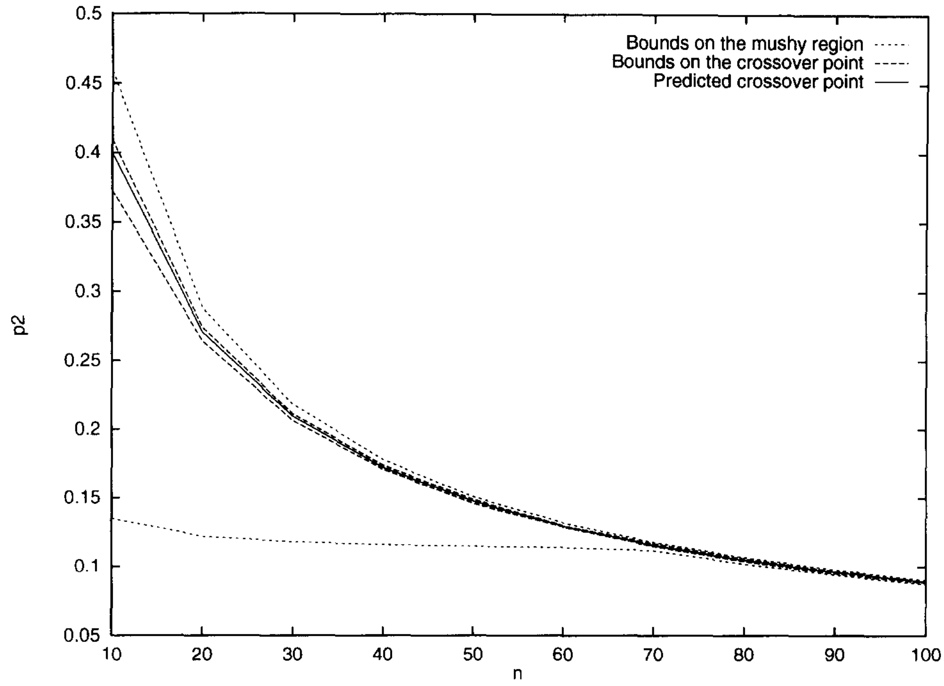

Fig. 7. Calculated phase transition bounds for $\langle n, n, 1\rangle$ problems.

Table 1
var(N) at the predicted crossover point for classes of CSPs with $p_1 = 1$

n *m*	10	20	30	40	50	60	70	80	90	100
10	3.34	10.8	31.1	88.3	251	723	2086	6036	17507	51029
20		3.32	5.57	9.44	16.0	27.4	46.8	80.2	137.4	235.9
30			3.06	4.34	6.18	8.79	12.5	17.9	25.5	36.4
40				2.96	3.85	5.02	6.53	8.50	11.1	14.4
50					2.91	3.59	4.43	5.47	6.74	8.32
60						2.88	3.43	4.08	4.86	5.78
70							2.86	3.32	3.85	4.46
80								2.85	3.24	3.69
90									2.83	3.18
100										2.82

As problems become sparser, Fig. 8 shows that there is an enormous increase in var(N) at $\hat{p}_{2\mathrm{crit}}$, culminating in the huge variances when $p_1 = 0.1$ already seen in Section 5. In general, for sparse CSPs in which m increases with n, evidence based on var(N) suggests that $\hat{p}_{2\mathrm{crit}}$ may eventually be a good predictor of the crossover point, but only for extremely large values of n: for $\langle n, n, 0.1\rangle$ problems, for instance, Fig. 8 shows that the variance at $\hat{p}_{2\mathrm{crit}}$ begins to decrease when $n = 60$, but clearly it will not reach the level seen for $p_1 = 1$ for a long time. Although experimental evidence suggests that

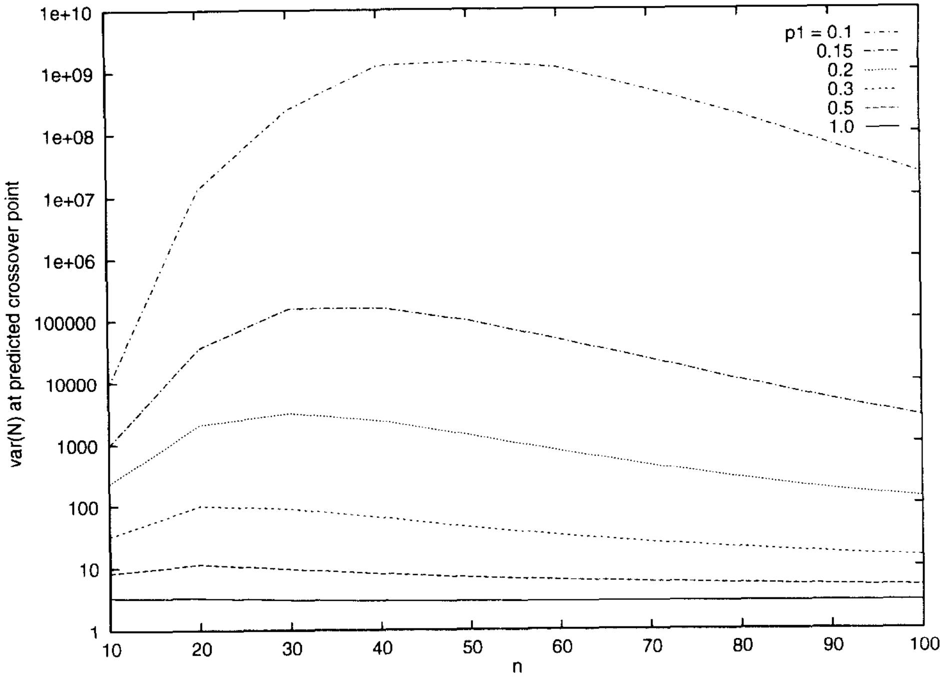

Fig. 8. var(N) at the predicted crossover point for $\langle n, n, p_1\rangle$ problems.

the phase transition does become sharper for $\langle n, n, 0.1\rangle$ problems as n increases, there is little evidence to support this from the variance; the calculated bounds on the mushy region are very far apart, although they appear to get closer as n increases. The situation is further complicated by the fact that for small sparse CSPs, different constraint graphs may have crossover points which differ significantly, as shown in Section 5. The variance of the number of solutions will also depend on the constraint graph, but var(N) for a particular constraint graph cannot easily be computed, even if it would be useful to do so.

Table 1 suggests that for some other classes of problems in which m increases with n, e.g. $\langle n, n/2, 1\rangle$, var(N) at $\hat{p}_{2\text{crit}}$ decreases as n increases. A similar pattern to the $\langle n, n, p_1\rangle$ problems might be expected in this case, i.e. for problems with high constraint density, $\hat{p}_{2\text{crit}}$ is a good predictor of the crossover point even for small values of n; as problems become sparse, however, this is only true for very large values of n.

Keeping m constant as n increases does not, asymptotically, make sense, since the minimum possible non-zero value of p_2 is $1/m^2$ and for any fixed p_2, $E(N) \to 0$ as $n \to \infty$, so that ultimately all such problems with $p_2 > 0$ are insoluble. However, for small n, the behaviour of this class of problem is of interest. If m is constant, Table 1 shows that var(N) at $\hat{p}_{2\text{crit}}$ increases as n increases: for $\langle n, 10, 1\rangle$, var(N) at $\hat{p}_{2\text{crit}}$ increases rather rapidly with n. From other investigations with different values of p_1, it appears to be a general rule that if m is kept constant, var(N) at $\hat{p}_{2\text{crit}}$ increases with n. This suggests that, except for small values of n, $\hat{p}_{2\text{crit}}$ gives a point in the insoluble

Table 2
$E(N)$ and $\mathrm{var}(N)$ for $\langle 200, 10, 1\rangle$ problems

p_2	$E(N)$	$\mathrm{var}(N)$
0.021	3.76×10^{16}	1.06×10^{35}
0.022	5.54×10^{7}	4.07×10^{19}
0.023	0.0798	8.98×10^{7}
0.024	1.13×10^{-10}	0.00127

region, rather than the crossover point. On the other hand, the bounds on the crossover point and the mushy region for $\langle n, 10, 1\rangle$ do appear to converge, though slowly, as n increases. For $n = 10$, they show that the crossover point occurs between 0.37 and 0.41 (empirically, $p_{2\text{crit}} = 0.4$, which is also the value given by (2)); when $n = 200$, the crossover point occurs between 0.01 and 0.03.

Since the calculated bounds on the crossover point, which are converging slowly, include $\hat{p}_{2\text{crit}}$, and on the other hand $\hat{p}_{2\text{crit}}$ appears to become a worse predictor of the crossover point, there is an apparent contradiction. This can perhaps be explained by considering a particular case in more detail. Table 2 shows $E(N)$ and $\mathrm{var}(N)$ for $\langle 200, 10, 1\rangle$ problems, increasing p_2 this time in steps of 0.001 (although this is unrealistic in terms of Model B). The calculated bounds on the crossover point are now 0.011 and 0.023; the predicted crossover point is at $p_2 = 0.0229$. Both $E(N)$ and $\mathrm{var}(N)$ are changing very rapidly at this point: it seems likely that the phase transition occurs over a very small range of values of p_2 for these problems, and that $\hat{p}_{2\text{crit}}$ lies outside the mushy region (as it does for the $\langle 30, 10, 0.1\rangle$ problems), where almost all problems have no solutions. In practice, given that with $m = 10$, p_2 cannot be varied in steps of less than 0.01 with Model B, it may be impossible to generate problems in the mushy region for these parameter values. For still larger values of n, the crossover point will in theory fall below 0.01, so that, as already pointed out, almost all problems with $p_2 > 0$ will be insoluble.

The class $\langle n, 10, 0.1\rangle$, an example of which was discussed in Section 5, is very badly-behaved: it combines the difficulties caused by sparse constraint graphs with those of the problems just discussed, in which m remains constant as n increases. The bounds on the crossover point only show that it lies between 0.61 and 0.8 when $n = 30$ (from Section 5, $p_{2\text{crit}} = 0.73$), and between 0.29 and 0.45 when $n = 80$. Together with the known inaccuracy of the predictor $\hat{p}_{2\text{crit}}$ when $n = 30$, this suggests that it will be difficult to predict the crossover point with any confidence for this class of problem, even for quite large values of n (and ignoring the complications caused by different constraint graphs). In the absence of experimental evidence about a class of CSPs of this type, there is little that can be said about the location of the crossover point, except that it lies within the broad limits given by the calculated bounds.

7. Practical applications

The ultimate aim in investigating phase transition phenomena is to be able to predict where an individual problem lies in relation to the phase transition, and hence to predict

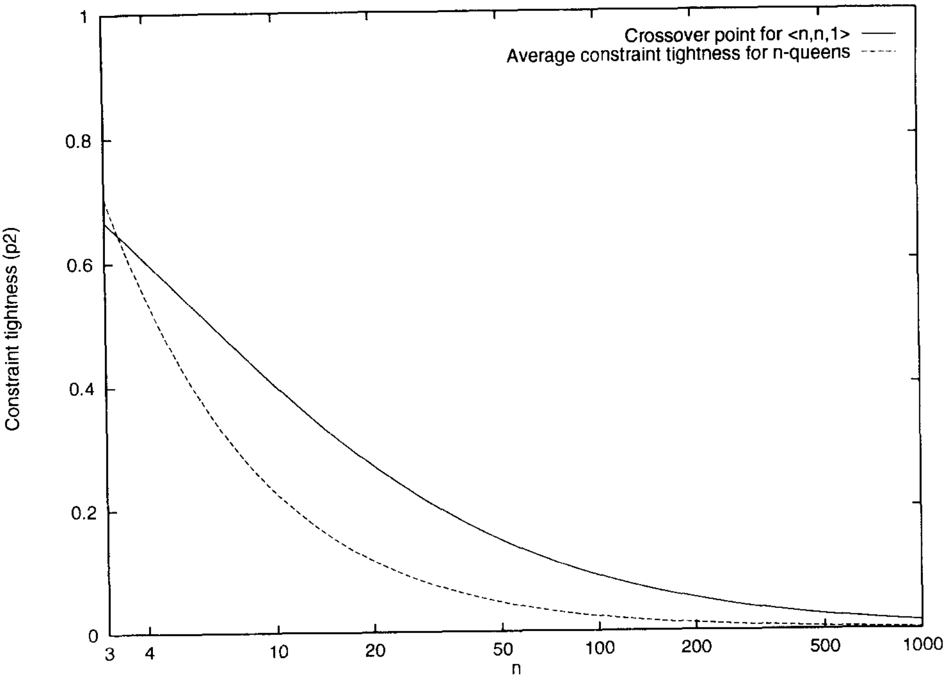

Fig. 9. Comparison of n-queens and $\langle n, n, 1\rangle$ problems.

whether it is almost certain to be insoluble (in which case there is probably no point in making the attempt), almost certain to be easily soluble, or (in the phase transition region) a hard problem which may or may not be soluble.

It has been shown in this paper that for many classes of randomly-generated binary CSPs, described by the parameters n, m, p_1 and p_2, the crossover point marking the transition from soluble to insoluble problems can be accurately estimated, by calculating the value of p_2 at which $E(N) = 1$, and it can be shown that the mushy region lies within a small range of values of p_2. Provided that a problem can be thought of as a random instance of one of these classes, it should be possible to make a prediction about whether or not it is likely to be soluble.

Unfortunately, a great many CSPs that arise in practice do not fit the model, for instance non-binary CSPs. It is also not obvious whether the kinds of constraint that arise in practice can be expected to behave like randomly-generated constraints. A great deal more work will be required in order to establish whether the work described here can be applied to real cases. One practical difficulty is that a large number of problems, both soluble and insoluble, will be required in order to assess the accuracy of the prediction.

A problem that appears frequently in the CSP literature is the n-queens problem,[8] partly because it does give an infinite number of instances (one for each n). It does

[8] The problem of placing n queens on an $n \times n$ chessboard in such a way that no queen can take any other.

also fit the model discussed in this paper reasonably well: the constraints are binary and every variable has the same number of values. The number of variables (the n queens) is equal to the number of values (the number of columns on the board), and every variable constrains every other variable, so that $p_1 = 1$. Every value for a variable conflicts with at most 3 values of any other variable, so the constraint tightness, p_2, is roughly $3/n$. More precisely, the constraint tightness varies to some extent, depending on the variables and values concerned: its average value can be calculated as $(7n-2)/3n^2$. Apart from the fact that p_2 varies, and the constraints are far from random, since they reflect the rules of chess, the n-queens problem can be seen as an instance of a $\langle n, n, 1, (7n-2)/3n^2\rangle$ problem. It has been demonstrated in the previous section that for the $\langle n, n, 1\rangle$ class, the crossover point occurs at $\hat{p}_{2\text{crit}}$ given by Eq. (2) and the phase transition is very abrupt, even for small n. Fig. 9 compares the average constraint tightness for n-queens with the predicted crossover point for $\langle n, n, 1\rangle$ problems. Problems falling below the solid curve are therefore predicted to be soluble; problems above it are predicted to be insoluble.

Since the n-queens problem is soluble for $n > 3$, and has an increasingly large number of solutions as n increases, it is gratifying that, as Fig. 9 shows, its behaviour agrees with the prediction given by $\hat{p}_{2\text{crit}}$. What is more interesting, however, is that for large n, Fig. 9 shows that almost all $\langle n, n, 1\rangle$ problems have no solutions; only problems with very loose constraints are soluble. To some extent, this counters one of the objections sometimes made to using n-queens as a benchmark problem, namely that it is unrepresentative because the constraints become looser as n gets larger (see Tsang [12], for example); Fig. 9 shows that if this were not the case, the problem would not remain soluble. However, the constraint tightness for the n-queens problem as a proportion of $\hat{p}_{2\text{crit}}$ does become smaller as n increases, so that it is an increasingly easy example of $\langle n, n, 1\rangle$.

8. Conclusions

It has been demonstrated that in randomly-generated binary CSPs defined by the parameters n, m, p_1 and p_2, as described in Section 2, there is a phase transition as the constraint tightness, p_2, increases, from a region in which almost all problems are soluble to a region in which almost all are insoluble. Between these two regions, in the mushy region, the average cost of finding a solution or proving that the problem is insoluble, is greatest. It is assumed that the peak in average cost occurs at the crossover point, $p_{2\text{crit}}$, where 50% of problems have solutions.

By considering a sample class of problems, $\langle 8, 10, 1\rangle$, it has been shown that the phase transition is also a transition from a partial search, which terminates as soon as the first solution is found, to a complete search, which is required to prove that there are no solutions. Hence, if all solutions are required, necessitating a complete search in all cases, the transition from solubility to insolubility does not correspond to a peak in average cost.

Experimental evidence and theory indicate that the crossover point occurs at the value of p_2 for which the expected number of solutions, $E(N)$, is 1, and that as the size of problems increases the phase transition should become increasingly abrupt, so that asymptotically there is an instantaneous phase transition. The accuracy of $\hat{p}_{2\text{crit}}$ as a predictor of $p_{2\text{crit}}$ has been investigated.

Although it has been demonstrated that $\hat{p}_{2\text{crit}}$ is an accurate estimate of $p_{2\text{crit}}$ for some classes of CSP, even when n is small, for instance for $\langle 8, 10, 1\rangle$, it is very inaccurate, and indeed does not even give a point in the mushy region, for some sparse problems when n is small. Detailed consideration of $\langle 30, 10, 0.1\rangle$ problems has shown that the phase transition also depends on the constraint graph for these problems, so that a more accurate predictor of the crossover point would have to be based on the constraint graph topology as well as on the other parameters. Although this is true in theory for any CSP with $p_1 < 1$, it is unlikely to be important unless the mushy region is wide (i.e. for small n and small p_1). For larger problems, the phase transition will happen sufficiently quickly that the effect of different constraint graphs will be insignificant. It was shown that for the $\langle 30, 10, 0.1\rangle$ problems, the reason for the poor performance of the $\hat{p}_{2\text{crit}}$ predictor is the very large variance in the number of solutions when $E(N) = 1$, so that almost all problems are insoluble, giving a point outside the mushy region, rather than the crossover point.

By calculating $\text{var}(N)$, the variance of the number of solutions, as well as $E(N)$, it is possible to derive bounds on the mushy region and the crossover point, and to determine whether $\hat{p}_{2\text{crit}}$ is likely to be an accurate estimate at the crossover point. Assuming that the trends seen in the calculated values continue as n increases, four classes of CSP have been identified:

- Problems with high constraint density and m increasing with n, of which $\langle n, n, 1\rangle$ is typical. Even for small n, $\hat{p}_{2\text{crit}}$ is an accurate estimate of the crossover point and there is an abrupt phase transition.
- Sparse problems in which m increases with n, e.g. $\langle n, n, 0.1\rangle$. For small values of n, the calculated bounds on the crossover region are very wide and $\text{var}(N)$ at $\hat{p}_{2\text{crit}}$ is extremely large, indicating that it is likely to be an over-estimate of the crossover point. There are indications that the situation improves, but only for very large values of n.
- Dense problems in which m is constant as n increases, e.g. $\langle n, 10, 1\rangle$. For large enough n, the crossover point will be below the smallest possible value of p_2 allowed by the model, so that almost all of these problems are insoluble. The calculated bounds on the crossover point appear to converge as n increases, but $\text{var}(N)$ at $\hat{p}_{2\text{crit}}$ is increasing, indicating that it is becoming more unreliable as an estimate of the crossover point. It is suggested that as the phase transition becomes increasingly abrupt, $\hat{p}_{2\text{crit}}$ may become numerically closer to the crossover point for these problems, although in the insoluble region.
- Sparse problems in which m is constant as n increases, e.g. $\langle n, 10, 0.1\rangle$. This class compounds the difficulties of the previous two classes. It is difficult to see how to locate the crossover point with any precision for these problems, other than by experimentation.

As discussed in the previous section, a great deal of further work is needed to see whether it is possible to apply these results to real, rather than randomly-generated, CSPs. If so, it might be possible to avoid wasting time trying to solve problems which can be predicted to be almost certainly insoluble. Alternatively, if a problem falls in the mushy region and so is likely to be hard, with a good chance of being insoluble, a small relaxation of the constraints would move it into the region where problems are almost certainly soluble and much easier to solve; the lower bound on the mushy region, given by the variance and expectation of the number of solutions, indicates by how much the constraints would need to be relaxed. Although the evidence provided by the n-queens problem is far from conclusive, it is at least an indication that this might be possible.

Appendix A

The derivation of $E(N)$ and $E(N^2)$ for problems generated according to Model B is as follows.

Let $c = \binom{n}{2} p_1$. Define the following indicator variables.

$$W_{ij} = \begin{cases} 1, & \text{if edge } \{i,j\} \text{ is present in the constraint graph,} \\ 0, & \text{otherwise,} \end{cases}$$

$$Z_{iy_i,jy_j} = \begin{cases} 1, & \text{if variable } i = y_i \text{ and variable } j = y_j \\ & \text{is an inconsistent assignment,} \\ 0, & \text{otherwise.} \end{cases}$$

Thus, in particular, $E(W_{ij}) = p_1$ and $E(Z_{iy_i,jy_j}) = p_2$. We will use the following notation for sums and products. Let

$$\sum_x = \sum_{x_1=1}^{m} \sum_{x_2=1}^{m} \cdots \sum_{x_n=1}^{m} \quad \text{and} \quad \prod_{i,j} = \prod_{i=1}^{n} \prod_{j=i+1}^{n},$$

with the understanding that different symbols replace all occurrences of the dummies. Thus, for example, $\prod_{r,s} = \prod_{r=1}^{n} \prod_{s=r+1}^{n}$. Then we have

$$N = \sum_x \prod_{i,j} (1 - W_{ij} Z_{ix_i,jx_j}). \tag{A.1}$$

To determine $E(N)$, note that there are m^n terms in the summation, each having the same expectation. Thus we only need to determine the expectation of the product for a fixed assignment x_i $(i = 1, 2, \ldots, n)$. However, only c terms in the product will be different from 1, those corresponding to $W_{ij} = 1$, and these will be independent with expectation $(1 - p_2)$ since they arise from different edges. Thus

$$E(N) = m^n (1 - p_2)^c,$$

as claimed. To determine $E(N^2)$, we have

$$\begin{aligned} N^2 &= \sum_x \sum_y \prod_{i,j} (1 - W_{ij} Z_{ix_i,jx_j}) \prod_{r,s} (1 - W_{rs} Z_{ry_r,sy_s}) \\ &= \sum_x \sum_y \prod_{i,j} (1 - W_{ij} Z_{ix_i,jx_j})(1 - W_{ij} Z_{iy_i,jy_j}). \end{aligned} \tag{A.2}$$

Consider the product in (A.2). Again only c terms in the product will be different from 1, those corresponding to $W_{ij} = 1$. If $W_{ij} = 1$, the term is $(1 - Z_{ix_i,jx_j})(1 - Z_{iy_i,jy_j})$, and these are independent for different edges $\{i,j\}$. If $x_i = y_i$ and $x_j = y_j$, then $(1 - Z_{ix_i,jx_j})(1 - Z_{iy_i,jy_j}) = (1 - Z_{ix_i,jx_j})$, which has expectation $(1 - p_2)$. Otherwise, if $x_i \neq y_i$ or $x_j \neq y_j$, then $E((1 - Z_{ix_i,jx_j})(1 - Z_{iy_i,jy_j}))$ is the probability q that two successive observations, drawn *without replacement* from a population of size m^2 containing m^2p_2 successes, are both failures. Clearly

$$q = \frac{(m^2 - m^2p_2)(m^2 - m^2p_2 - 1)}{m^2(m^2 - 1)} = (1 - p_2)\left(1 - \frac{m^2p_2}{m^2 - 1}\right).$$

Thus the expectation of the product is determined only by the number of edges $\{i,j\}$ such that $x_i = y_i$ and $x_j = y_j$. Thus, let $S = \{i : x_i = y_i\}$. The number of ways of choosing $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ so that exactly s pairs match, i.e. $x_i = y_i$ for exactly s values of i, is

$$\binom{n}{s} m^n (m - 1)^{n-s}.$$

Hence, this is the number of terms in the summation of (A.2) with $|S| = s$, for $s = 0, 1, \ldots, n$. Now for given S, the probability that we choose t edges with both vertices in S is the probability of t successes in samples of size c drawn without replacement from a population of $\binom{n}{2}$ containing $\binom{s}{2}$ successes. This has the hypergeometric probability

$$\frac{\binom{\binom{s}{2}}{t}\binom{\binom{n}{2}-\binom{s}{2}}{c-t}}{\binom{\binom{n}{2}}{c}},$$

for $t = 0, 1, \ldots, c$. Given s, the product in (A.2) has expectation $(1 - p_2)^t q^{c-t}$ by the argument above. Putting all this together,

$$E(N^2) = \sum_{s=0}^{n} \binom{n}{s} m^n (m - 1)^{n-s} \sum_{t=0}^{c} \frac{\binom{\binom{s}{2}}{t}\binom{\binom{n}{2}-\binom{s}{2}}{c-t}}{\binom{\binom{n}{2}}{c}} (1 - p_2)^t q^{c-t},$$

and so, as claimed:

$$E(N^2) =$$

$$m^n \sum_{s=0}^{n} \binom{n}{s} (m - 1)^{n-s} \sum_{t=0}^{c} \frac{\binom{\binom{s}{2}}{t}\binom{\binom{n}{2}-\binom{s}{2}}{c-t}}{\binom{\binom{n}{2}}{c}} (1 - p_2)^c \left(1 - \frac{m^2p_2}{m^2 - 1}\right)^{c-t}. \quad \text{(A.3)}$$

A similar analysis can be done for Model A. For the expectation, since all terms in (A.1) are independent, we obtain

$$E(N) = m^n (1 - p_1p_2)^{\binom{n}{2}}.$$

For the variance computation, however, $c = t_1 + t_2$ is now a random variable, where t_1 is the number of edges with both endpoints in S. We now obtain

$$q = E((1 - p_2)^{t_1}(1 - p_2)^{t_2}) = E((1 - p_2)^{t_1})E((1 - p_2)^{t_2}),$$

since t_1, t_2 are independent. But these are binomial random variables with parameters $\binom{s}{2}$, p_1 and $\binom{n}{2} - \binom{s}{2}$, p_1 respectively. Since for a binomial variable t with parameters ν, p we have $E(\lambda^t) = (1 - p(1 - \lambda))^{\nu}$, we obtain

$$\begin{aligned} q &= (1 - p_1p_2)^{\binom{s}{2}} (1 - p_1(1 - (1 - p_2)^2))^{\binom{n}{2}-\binom{s}{2}} \\ &= (1 - p_1p_2)^{\binom{s}{2}} (1 - p_1p_2(2 - p_2))^{\binom{n}{2}-\binom{s}{2}}. \end{aligned}$$

This gives

$$E(N^2) = m^n \sum_{s=0}^{n} \binom{n}{s} (m-1)^{n-s} (1 - p_1p_2)^{\binom{s}{2}} (1 - p_1p_2(2 - p_2))^{\binom{n}{2}-\binom{s}{2}}. \quad (A.4)$$

A detailed asymptotic analysis of these variance formulas appears tricky, so we will consider only two special cases with $m = n$.

First, consider the case of Model B and $p_1 = 1$. Then $c = \binom{n}{2}$ and if $E(N) \approx 1$, we require $p_2 \approx 2\log n/(n-1)$. Since then $m^2p_2/(m^2 - 1) = p_2(1 + \mathrm{O}(1/n^2))$, we may approximate $(1 - m^2p_2/(m^2 - 1))$ by $(1 - p_2)$. The inner sum in the expression for $E(N^2)$ in (A.3) reduces to a single term, with $t = \binom{s}{2}$, giving

$$\begin{aligned} E(N^2) &\approx \sum_{s=0}^{n} \binom{n}{s} (n-1)^{n-s} (1 - p_2)^{\binom{n}{2}-\binom{s}{2}} \\ &\approx \mathrm{e}^{-1} \sum_{s=0}^{n} \binom{n}{s} (n-1)^{-s} (1 - p_2)^{-\binom{s}{2}} \\ &\approx \mathrm{e}^{-1} \sum_{s=0}^{n} \binom{n}{s} (n-1)^{-s} n^{s(s-1)/(n-1)} \\ &\approx \mathrm{e}^{-1} \sum_{s=0}^{n} \binom{n}{s} \mathrm{e}^{s/n} n^{s(n-s)/(n-1)}, \end{aligned}$$

using the values of $E(N)$ and p_2. Examination of the final sum shows that the terms are negligible unless s is near zero or s is near n. Then

$$E(N^2) \approx (1 + \mathrm{e}^{-1}) \sum_{s=0}^{\infty} 1/s! = (1 + \mathrm{e}),$$

giving $\operatorname{var}(N) \approx \mathrm{e}$.

By contrast, in Model A, for any constant p_1 we require

$$p_2 \approx 2\log n/(n-1)p_1.$$

Taking only the term of $E(N^2)$ for $s = 0$, from (A.4), gives

$$n^n(n-1)^n(1 - p_1p_2(2 - p_2))^{\binom{n}{2}} \geqslant \frac{1}{4} \left(1 + \frac{p_1p_2^2(1 - p_1)}{(1 - p_1p_2)^2}\right)^{\binom{n}{2}},$$

using the value of $E(N)$. This is at least

$$\frac{1}{4}\left(1+\frac{p_2^2}{4}\right)^{\binom{n}{2}} \geqslant \frac{1}{4}\left(1+\frac{\log^2 n}{4p_1^2}\right),$$

for large n. This clearly diverges to infinity. Thus in this case the variance is unbounded, even when $p_1 = 1$. Thus the two models are not equivalent.

References

[1] B. Bollobás, *Random Graphs* (Academic Press, New York, 1985).

[2] P. Cheeseman, B. Kanefsky and W. Taylor, Where the Really Hard Problems are, in: *Proceedings IJCAI-91*, Sydney, Australia (1991) 331–337.

[3] J.M. Crawford and L.D. Auton, Experimental results on the crossover point in satisfiability problems, in: *Proceedings AAAI-93*, Washington, DC (1993) 21–27.

[4] R. Dechter and J. Pearl, Network-based heuristics for constraint-satisfaction problems, *Artif. Intell.* **34** (1988) 1–38.

[5] R. Haralick and G. Elliott, Increasing tree search efficiency for constraint satisfaction problems, *Artif. Intell.* **14** (1980) 263–313.

[6] S. Kirkpatrick and B. Selman, Critical behaviour in the satisfiability of random Boolean expressions, *Science* **264** (1994) 1297–1301.

[7] D.G. Mitchell, B. Selman and H.J. Levesque, Hard and easy distributions of SAT problems, in: *Proceedings AAAI-92*, San Jose, CA (1992) 459–465.

[8] E.M. Palmer, *Graphical Evolution* (Wiley, New York, 1985).

[9] P. Prosser, Hybrid algorithms for the constraint satisfaction problem, *Comput. Intell.* **9** (3) (1993) 268–299.

[10] P. Prosser, Binary constraint satisfaction problems: some are harder than others, in: A. Cohn, ed., *Proceedings ECAI-94* (Wiley, New York, 1994) 95–99.

[11] P. Prosser, An empirical study of phase transitions in binary constraint satisfaction problems, *Artif. Intell.* **81** (1996) 81–109 (this volume).

[12] E. Tsang, *Foundations of Constraint Satisfaction* (Academic Press, New York, 1993).

[13] C. Williams and T. Hogg, Using deep structure to locate hard problems, in: *Proceedings AAAI-92*, San Jose, CA (1992) 472–477.

[14] C. Williams and T. Hogg, Extending deep structure, in: *Proceedings AAAI-93*, Washington, DC (1993) 152–158.

[15] C. Williams and T. Hogg, Exploiting the deep structure of constraint problems, *Artif. Intell.* **70** (1994) 73–117.