Clustering at the Phase Transition

Andrew J. Parkes
CIS Dept. and CIRL
1269 University of Oregon
Eugene, OR 97403-1269
parkes@cirl.uoregon.edu

Abstract

Many problem ensembles exhibit a phase transition that is associated with a large peak in the average cost of solving the problem instances. However, this peak is not necessarily due to a lack of solutions: indeed the average number of solutions is typically exponentially large. Here, we study this situation within the context of the satisfiability transition in Random 3SAT. We find that a significant subclass of instances emerges as we cross the phase transition. These instances are characterized by having about 85–95% of their variables occurring in unary prime implicates (UPIs), with their remaining variables being subject to few constraints. In such instances the models are not randomly distributed but all lie in a cluster that is exponentially large, but still admits a simple description. Studying the effect of UPIs on the local search algorithm WSAT shows that these "single-cluster" instances are harder to solve, and we relate their appearance at the phase transition to the peak in search cost.

Introduction

A phase transition in a physical many-body system is the abrupt change of its macroscopic properties at certain values of the defining parameters. The most familiar example is the water/ice transition in which the fluidity changes abruptly at particular temperatures and pressures. Phase transitions also occur in systems associated with computer science: work in this area started with the remarkable observation (Erdös & Rényi 1960) that thresholds in properties such as connectivity emerge in large random graphs. Recently, phase transitions have been studied in constraint satisfaction, e.g. see (Cheeseman, Kanefsky, & Taylor 1991; Mitchell, Selman, & Levesque 1992; Williams & Hogg 1993; Kirkpatrick & Selman 1994; Smith 1994). In all these cases we have "control" parameters defining the system, a method to generate an ensemble of problem instances given values of such parameters, and some property A whose existence or not we wish to determine for each problem instance. When the problems are sufficiently large, then small variations in the control parameters can mean that the ensembles quickly change from having almost all instances satisfying the property A, to having almost no instances satisfying A.[1]

It can also be that we need to search in order to determine whether or not property A is satisfied (e.g. determining satisfiability for SAT problems) and then the phase transition is typically associated with a peak in the cost of such a search. It becomes easier to find either a witness to A, or a proof of its non-existence, as we move away from the transition. This peak has (at least) two impacts on artificial intelligence. Firstly, even in real problems we might well see phase transitions along with the peak in search cost (Huberman & Hogg 1987). Secondly, the development of search algorithms has been plagued by the lack of hard test problems. If we use real problems then the test set is likely to be small and we have the risk of over-fitting. To make the algorithms more robust it is useful to be able to generate a large number of test instances. However, initial attempts at artificial generation gave unrealistically easy instances. This situation was relieved by the discovery that phase transitions can be used as sources of hard artificial problems (Mitchell, Selman, & Levesque 1992). This suggests that deeper study of such transitions is relevant to AI.

If we refer to a witness to the property A as a model, then in many random systems it is possible to determine the average number of models in terms of the control parameters. At the phase transition the average number of models is typically *exponential* in the problem size, and this immediately raises the questions:

- how are the many models distributed amongst instances of the ensemble?
- for a particular instance how are the models (if any) distributed in the space of assignments?

Furthermore, even if we restrict ourselves to satisfiable instances and use local search, then the search

[1] It is often convenient to refer to the crossover point: parameter values for which 50% of the instances satisfy A.

cost still seems to peak at the phase transition (Hogg & Williams 1994; Clark *et al.* 1996). This is somewhat counter-intuitive because the average number of models per instance does not peak (or even seem to be special in any way), yet we might well expect that the search cost is related to numbers of models.

In this paper we study such issues in the context of the well-studied satisfiability transition in Random 3SAT (Kirkpatrick & Selman 1994; Crawford & Auton 1996; Schrag & Crawford 1996). Our aim is to provide a finer description of the transition than the coarse description as a boundary between satisfiable and unsatisfiable phases. However, the exponential number of models also makes it difficult to study them directly. Instead we work indirectly by looking at the implicates of the theory and in particular at the distributions associated with the unary prime implicates[2] (UPIs). Our main result is that there is indeed a finer structure observable in the UPI–distributions. As we cross into the unsatisfiable phase then there emerges a large distinct subclass of instances. The models in these instances are not randomly distributed, but all lie in a single exponentially large cluster, which moreover admits a short and simple description. These "single cluster instances" are harder to solve by local search, and their emergence at the phase transition seems to be linked to the peak in search cost.

Random 3SAT background

By Random 3SAT we mean the uniform fixed-length clause distribution (Mitchell, Selman, & Levesque 1992). Ensembles are parameterized by (n, c) where n is the total number of variables, and c is the number of clauses. The c clauses are generated independently by randomly picking 3 variables and negating each variable with probability of 50%. The clause/variable ratio, $\alpha = c/n$, is used as the control parameter.

There are 2^n possible variable assignments, and any one clause is consistent with 7/8 of these. The average number of models, M, per problem instance is then (e. g. (Williams & Hogg 1993))

$$M = 2^n \left(\frac{7}{8}\right)^c = 2^{n(1\,-\,\alpha/5.19)} \qquad (1)$$

Any region with $M < 1$ must contain unsatisfiable instances, so assuming the existence of a satisfiability phase transition forces it to occur at $\alpha \leq 5.19$. Experiments show that the phase transition actually occurs at $\alpha \approx 4.26$ (Crawford & Auton 1996) giving an average of $2^{0.18n}$ models per instance. There are arguments that come quite close to this empirical value (Williams & Hogg 1993; Smith 1994), but so far there is no derivation of the *exact* transition point. In contrast, for random 2SAT the position of the phase transition is known exactly, but is not so interesting because satisfiability is then in **P**. Actually, there appears to be an intriguing link between exact results and the complexity class of the decision problem in that exact results are associated with **P**. The link is often direct, as for 2SAT, but might also be indirect, e. g. the threshold for the existence of Hamiltonian cycles (which is NP-complete) is known exactly in random graphs. However, this threshold "piggy-backs" the transition for 2-connectedness, which is in **P**. As soon as the graph is 2-connected then it almost surely has a Hamiltonian cycle (Bollobás 1985). I am not aware of exact threshold results not of one of these types.

[2] An implicate is entailed *by* the theory. A prime implicate is one that is not subsumed by another implicate. A unary implicate is just a literal. A binary implicate, for our purposes, is a disjunction of two literals, i. e. a clause of length two.

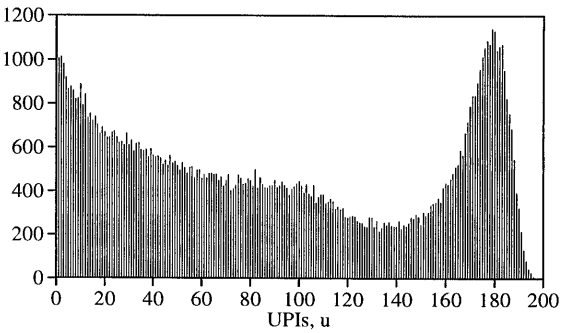

Figure 1: Histogram giving the numbers of instances having u UPIs as a function of u. Obtained from 10^5 satisfiable instances at (n,c)=(200,854).

To remove unsatisfiable instances from the ensemble, and also to find the prime implicates we used NTAB, a variant of TABLEAU, (Crawford & Auton 1996; Schrag & Crawford 1996). Such systematic searches are computationally expensive and provide the main limit on the range of accessible values of n. Once we have the UPIs for an instance then we can also generate the "residual" theory. This is a mixed 2SAT/3SAT theory over all the residual, or free, variables (those variables not contained in UPIs). The residual clauses are obtained by using the UPIs to simplify each of the original clauses. If an instance has u UPIs then the residual theory has $f = n - u$ residual variables, and, by definition, must be satisfiable and have no UPIs itself.

Since some of the fastest known algorithms are local search algorithms we also study the performance of WSAT (Selman, Kautz, & Cohen 1994) on the satisfiable instances.

Results for Random 3SAT

In this section we study the "UPI–distribution" as a function of (n, c). After generating instances, we remove those that are unsatisfiable, and for each of the others we find the number of UPIs, u, that it possesses.

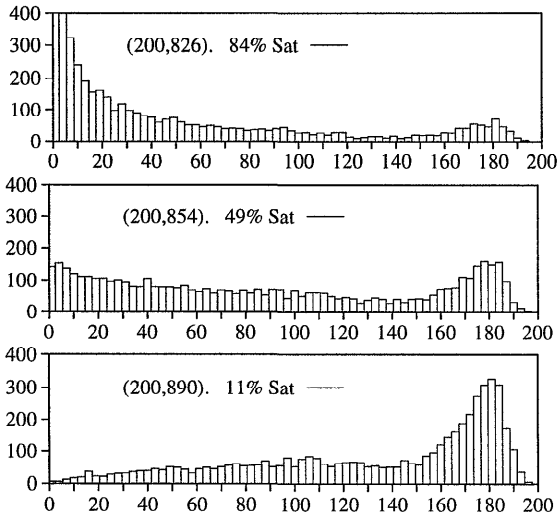

Figure 2: Histograms of the UPI–distribution obtained while traversing the phase transition with n fixed at 200. The legend "x% Sat" means that, at the given (n, c) values, x% of the total ensemble is satisfiable. Each histogram is derived from 5000 satisfiable instances.

Counting the number of instances that occur in given ranges of u gives the desired distribution.

Figure 1 gives the UPI–distribution found close to the crossover point for $n = 200$. Figure 2 gives a sample of the results obtained from a "slice" through the phase transition, i. e. varying c at fixed n.

At small c most instances have few UPIs, but the average number of UPIs increases dramatically as we increase c. A peak emerges in the region $170 < u < 190$, and for reasons to be explained later we will call this the "single cluster peak". Conversely, the region $120 < u < 170$ remains consistently underpopulated. Perhaps, figure 2 might best be described as a "sloshing of the UPIs".

We should check that this pattern persists as we increase n, and for this we use the UPI density u/n. Figure 3 shows the effect of increasing n while remaining at the crossover. The distribution pattern seems to persist (except for smaller values of u/n, with which we will not be concerned here). In particular, the position (and indeed existence) of the single-cluster peak is stable at $u/n \approx 0.9$.

Interpretation of the UPI slosh

In random graph theory it can be helpful to study the evolution of ensemble properties as we add more edges (Bollobás 1985). Here we make some simple arguments in order to study the evolution of the UPI-distribution as we add more clauses. In particular, we propose a partial explanation of the "slosh" of the UPIs as observed in figure 2.

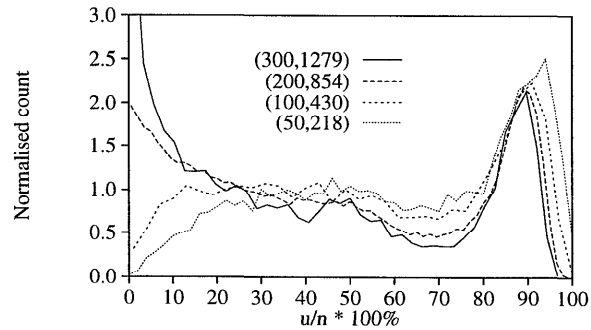

Figure 3: Histogram of UPI–distributions. Here the x-axis is percentage UPIs. The y-axis is the frequency of instances normalized so that a flat distribution would give 1. The (n, c) values are chosen to remain at the crossover point as we increase n.

Consider the addition of a single new and randomly selected clause to an existing instance. We want to find the number of new UPIs created as a result of this new clause, and in particular to find how the number of new UPIs varies as a function of the numbers of UPIs and number of binary prime implicates (BPIs) b already present in the instance. We will use figure 4 in order to illustrate the arguments for the crossover point at n=100. It is convenient to define $\lambda = f/n \equiv (1 - u/n)$. Since the new clause is independent of the existing UPIs we have that each literal of the new clause remains free with probability λ, otherwise the UPIs force it true with probability $(1 - \lambda)/2$ or false with probability $(1 - \lambda)/2$. It is then straightforward to find the probabilities of the various fates of the new clause under the existing UPIs. For example, there is a probability of $((1 - \lambda)/2)^3$ that all three literals of the new clause are forced to false rendering the theory unsatisfiable. We must exclude this case (as we only consider satisfiable instances) and then the probability of the new clause reducing to a unary clause, and so directly giving rise to a new UPI is

$$\text{Prob(Unit Clause)} = P(\lambda) = \frac{3\lambda(\frac{1-\lambda}{2})^2}{1 - \frac{1}{8}(1-\lambda)^3} \qquad (2)$$

An example of this function is given in figure 4c.

However, if we happen to create a new UPI then it can give rise to more UPIs by resolution against the b BPIs already present in the theory. Again, since the new clause (and hence initial new UPI) was independent of the existing BPIs it follows that the expected number of BPIs that contain the negation of the new UPI is just $(2b)/(2f)$. So, after a single pass through the BPIs, we will have a total of $(1 + b/f)P(\lambda)$ new UPIs. An example is given in figure 4d based on the empirical data for b.

These latest UPIs can again propagate to cause even more UPIs, but further calculation is hampered by the

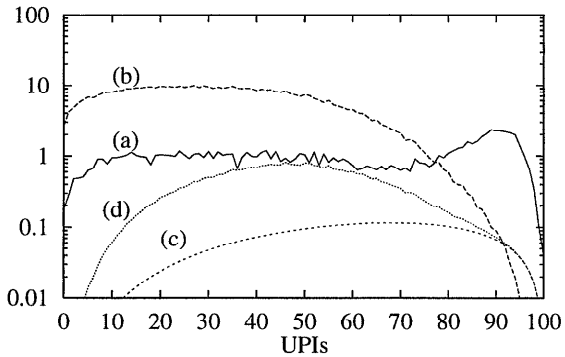

Figure 4: For a sample of 10^4 satisfiable instances at (100,430) we give, as a function of the number of UPIs: (a) the empirical (normalized) frequency of that number of UPIs, (b) the empirical average number of BPIs per free variable, (c) $P(\lambda)$, an estimate of the number of new UPIs directly created from one new clause. (d) as for (c) but also including the effect of propagating the new UPIs *once* through the BPIs.

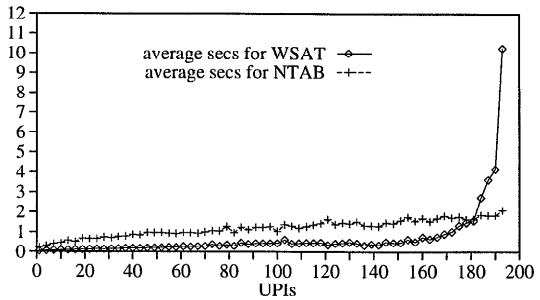

Figure 5: Mean times for NTAB and WSAT to find a solution as a function of the number of UPIs at $(n, c) = (200, 854)$. Based on 10^4 instances. Points above $u = 194$ are omitted due to insufficient instances.

fact that the latest UPIs need not be be independent of the existing theory. Even with this crude approximation we see from figure 4d that that the effect of a new clause is smallest when u is close to either 0 or n, and largest in the middle of the range.

This allows us to build a picture of the typical evolution of an instance starting in the underconstrained region and adding new clauses while ensuring that the theory does not become unsatisfiable. Being initially underconstrained the instance will start with a small value of u. At first the growth of u will be slow because of the small chance of new clauses generating new UPIs, however once u starts to grow then the peak observed in figure 4d suggests that the growth of u will be rapid until u reaches $u \approx 0.8n$. At this point growth of u will again slow down. This provides a reasonable "first-order" explanation of the deficit of instances in the middle and the accumulation of instances into the single-cluster peak.

Clusters

We now look at the nature of instances in the peak observed in the UPI distribution at $u/n \approx 0.9$. The most important property we find is that for instances in the peak, the residual theories (after imposing the UPIs) are very underconstrained, with few clauses per residual variable. This is supported by figure 4b – in the region $u > 0.8n$ there are few BPIs per residual variable. We have also confirmed this by other methods such as direct model counts in the peak region. Hence, these instances have no models outside of the region of the assignment space compatible with the UPIs, but inside this region there are few constraints and the density of models is relatively high: we call such a region a cluster. This cluster of models is compactly described in terms of an assignment to about 90% of the variables and a small number of constraints on the remaining variables. Note that the cluster is small compared to the total assignment space but still exponentially large in n.

Since the peak contains such "single-cluster instances" then it is reasonable to ask what would happen if we had two such clusters. For simplicity, suppose the clusters were independent, each defined by assignments to a random set of $0.9n$ variables, and with empty residual theories. In this case we only obtain a UPI when the two defining assignments both contain the variable, and also agree on its assignment. Since only 0.9^2n variables occur in both assignments we can expect to get about $0.4n$ UPIs. Adding more clusters would further reduce the number of UPIs, and so we would expect to have few instances with u between these single and double cluster values. Although the assumption of independence of the clusters is probably too strong, the jump of $u \approx 0.9n$ for single cluster instances down to $u \approx 0.4n$ for the hypothesised 2-cluster instances would explain the deficit of instances in the region 0.6–0.8 for u/n seen in figure 1. Also, figure 1 shows some indications of a secondary weak underlying peak at $u/n \approx 0.5$. However, it is clear that further work would be needed to investigate the accuracy of such a multi-cluster interpretation.

UPIs and search costs

We now study the effects of the number of UPIs on search cost. Firstly we briefly compare systematic search using NTAB and local search using WSAT. We give results for n=200 (for WSAT we use Maxflips=20000 and noise of 0.5). Figure 5 shows that for NTAB the mean runtime increases fairly smoothly with u, and that over the majority of the range WSAT is faster than NTAB. However, WSAT is badly affected by the single-cluster instances: WSAT spends 70% of its time on the the instances with $\geq$ 75%

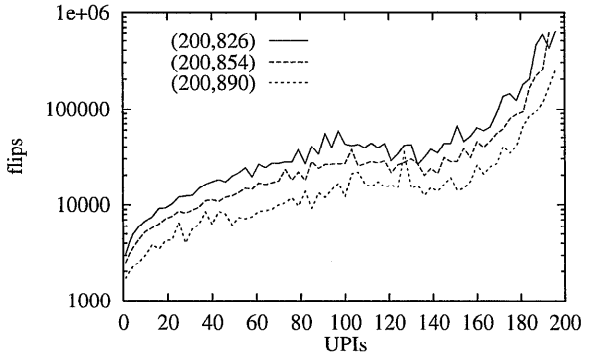

Figure 6: Average flips taken by WSAT plotted against the number of UPIs for various values of c at n=200.

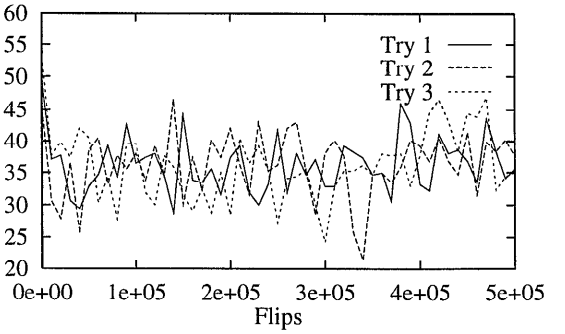

Figure 7: Percentage of UPIs violated against flips for 3 independent tries of WSAT on a hard instance at (350,1491), and having 326 UPIs.

UPIs, although they form only 27% of the sample. These effects occur for the median as well as the mean, and so we think it is unlikely that they are similar to the extra peaks seen in (Hogg & Williams 1994; Gent & Walsh 1994).

While NTAB is less sensitive to the UPIs it also scales much worse than WSAT (e. g. (Parkes & Walser 1996)) and so is not the best algorithm for the satisfiable instances (though the results do suggest that it might still remain competitive for those instances that have a lot of UPIs and so are very close to unsatisfiability). Hence, we now restrict ourselves to WSAT.

In figure 6 we look again at the slice through the transition region first considered in figure 2. Across the whole u range it shows that, in the phase transition region, if we fix u, while increasing c, then the theory gets easier. Presumably the extra clauses are helping to direct WSAT towards solutions. Initially this might seem inconsistent with the overall peak of hardness being at crossover point at (n,c)=(200,854). However, we have the competing effect that as we traverse the phase transition region then many instances have a rapid increase in their values for u and so will become much harder. Eventually, the average hardness will be dominated by instances already lying in the single-cluster peak and then the average u does not increase much further, and we will again see the drop in hardness observed at fixed u values. Similar effects are seen with the balance between satisfiable and unsatisfiable instances themselves (Figure 3 of (Mitchell, Selman, & Levesque 1992)).

Finally we return to the effects of new clauses. Suppose we have a cluster defined by $0.9n$ variables, then there is a probability of at least $0.9^3/8 \approx 0.1$ that a new clause will be violated by all the models in the cluster, converting the cluster into what we call a "failed cluster" (a region much like a cluster but containing no models). Given an instance with two clusters then it is quite possible that a single clause could be responsible for reducing it to a single cluster instance. The impact on search would arise from the fact that the "failed cluster" was initially exponentially large but was all removed by a single clause – local search is then likely to see it as a very attractive region that contains many assignments with just a single violated clause. The resulting exponentially large, but none-the-less false, minimum could easily impede progress.

In figure 7 we look at just one of the hardest single-cluster instances that we found at the crossover at $n = 350$. The progress of WSAT as a function of flips made is measured by the percentage of UPIs correctly matched by the variable assignment. The independent tries consistently settle into a region with many incorrect UPIs (many of the assignments also violate just one clause). The tries vary over the range 30-40% i. e. 10% of the variables which also matches the expected size of the clusters and failed clusters. Hence, this is consistent with the instance having a failed cluster that traps WSAT. Of course, this is just a partial study of just one instance and much more work would be needed to confirm this picture of failed clusters. If true then the exponential size of such failed clusters would also probably adversely affect techniques such as tabu lists that are usually used to escape such false minima. However, the shortness of the description of the clusters (just an assignment to about 90% of the variables) does offer the hope that if it were possible to find such descriptions then they could be used to help escape from the failed clusters.

Related work

The effect of the number of models on the cost of local search has been studied in (Clark *et al.* 1996). Since instances with large u generally have fewer models our observations on the effect of u are consistent with their confirmation of the expectation that instances with fewer models are harder to solve. The results based on model counting do not seem to show the same clear structure and evolution that we have seen for the UPI-counts. (On the other hand (Clark *et al.* 1996) did not

restrict themselves to Random 3SAT.)

A measure of clause imbalance, Δ, has been suggested as a useful parameter at the phase transition (Sandholm 1996). However, Δ itself does not undergo a transition, and is a refinement to α. In contrast, u/n changes abruptly from being nearly zero to being nearly one, and is a refinement to the satisfiability.

Properties of counts of prime implicates are also studied in (Schrag & Crawford 1996), but by taking averages over instances at a given value of (n,c) rather than looking at the distributions within the ensemble. However, they cover a wider range of parameter values, and also look at longer implicates.

Conclusions

The transition region has been described as being "mushy" (Smith 1994). We have seen that it can also be described as "slushy": consisting of a mix of frozen instances and fluid instances. The frozen instances have a lot of UPIs but their residual theory is underconstrained: hence all of their models lie in a single cluster of exponential size but very compact description. The fluid instances have few UPIs. As we traverse the phase transition then more instances freeze. The transition for individual instances is rather fast because instances with a medium number of UPIs are relatively unstable against the addition of extra clauses. Note that we do *not* regard the abrupt change in the number of UPIs as a different phase transition. Instead our view is that there is one transition with many different aspects, of which the changes in satisfiability and UPIs are just the most apparent.

The single-cluster instances have a large and deleterious effect on the local search algorithm WSAT, and often come to dominate the runtimes. It is likely that their emergence at the phase transition is largely responsible for the cost of local search peaking in this region. In contrast, NTAB is very sensitive to the unsatisfiability aspects and less affected by the UPIs. (Presumably, the many aspects of the phase transition mean that every algorithm meets at least some problem, though not necessarily always the same problem.)

We also saw some (weak) evidence for instances with two clusters, or one cluster along with a failed cluster. More work would be needed to see to what extent a multi-cluster description is useful. However, the compactness of the cluster description suggests that it might be of use for knowledge compilation, or for techniques to help local search avoid exponentially large, but false, minima.

Although we have only considered random 3SAT, it could be interesting to make similar investigations for random CSP problems. For example, one could look at how the κ parameter of (Gent *et al.* 1996) (itself directly related to the average number of models) is related to the position (or even existence) of the single-cluster peak.

Acknowledgments

I am indebted to James Crawford, Joachim Walser, and all the members of CIRL for many helpful comments and discussions. This work has been supported by ARPA/Rome Labs under contracts F30602-93-C-0031 and F30602-95-1-0023.

References

Bollobás, B. 1985. *Random graphs.* Academic Press, London.

Cheeseman, P.; Kanefsky, B.; and Taylor, W. M. 1991. Where the really hard problems are. *Proceedings IJCAI-91* 1.

Clark, D.; Frank, J.; Gent, I.; MacIntyre, E.; Tomov, N.; and Walsh, T. 1996. Local search and the number of solutions. In *Proceedings CP-96.*

Crawford, J. M., and Auton, L. 1996. Experimental results on the crossover point in random 3-SAT. *Artificial Intelligence* 81:31–57.

Erdös, P., and Rényi, A. 1960. *Publ. Math. Inst. Hung. Acad. Sci.* 5(7).

Gent, I., and Walsh, T. 1994. Easy Problems are sometimes Hard. *Artificial Intelligence* 70:335–345.

Gent, I. P.; MacIntyre, E.; Prosser, P.; and Walsh, T. 1996. The constrainedness of search. In *Proceedings AAAI-96*, 246–252.

Hogg, T., and Williams, C. P. 1994. The hardest constraint problems: a double phase transition. *Artificial Intelligence* 69:359–377.

Huberman, B. A., and Hogg, T. 1987. Phase transitions in artificial intelligence systems. *Artificial Intelligence* 33:155–171.

Kirkpatrick, S., and Selman, B. 1994. Critical behavior in the satisfiability of random boolean expressions. *Science* 264:1297–1301.

Mitchell, D.; Selman, B.; and Levesque, H. 1992. Hard and Easy Distributions of SAT problems. In *Proceedings AAAI-92*, 459–465.

Parkes, A. J., and Walser, J. P. 1996. Tuning Local Search for Satisfiability Testing. In *Proceedings AAAI-96*, 356–362.

Sandholm, T. W. 1996. A Second Order Parameter for 3SAT. In *Proceedings AAAI-96*, 259.

Schrag, R., and Crawford, J. M. 1996. Implicates and Prime Implicates in Random 3SAT. *Artificial Intelligence* 88:199–222.

Selman, B.; Kautz, H. A.; and Cohen, B. 1994. Noise strategies for improving local search. In *Proceedings AAAI-94*, 337–343.

Smith, B. M. 1994. Phase Transition and the Mushy Region in Constraint Satisfaction Problems. In *Proceedings ECAI-94*, 100–104.

Williams, C. P., and Hogg, T. 1993. Extending Deep Structure. In *Proceedings AAAI-93*, 152–157.