Random Horn Formulas and Propagation Connectivity for Directed Hypergraphs[†]

Robert H. Sloan[1] Despina Stasi[1] György Turán[1,2]

[1] *University of Illinois at Chicago*
[2] *Hungarian Academy of Sciences and University of Szeged, Research Group on Artificial Intelligence*

received 17th October 2010, *revised 21st March 2012*, *accepted 1st August 2012*.

We consider the property that in a random definite Horn formula of size-3 clauses over n variables, where every such clause is included with probability p, there is a pair of variables for which forward chaining produces all other variables. We show that with high probability the property does not hold for $p \leq 1/(11n \ln n)$, and does hold for $p \geq (5 \ln \ln n)/(n \ln n)$.

Keywords: random Horn formulas, propagation connectivity

1 Introduction

Horn formulas are a subclass of CNF expressions, where every clause contains at most one unnegated variable. This class is tractable in the sense that many problems that are hard for CNF expressions in general are polynomially solvable for Horn formulas (such as satisfiability and equivalence). It is partly for this reason that Horn formulas are of basic importance in artificial intelligence and other areas. Random Horn formulas have been studied in [DBC01, DV06, Ist02, LMST09, MIDV07].

A Horn formula is *definite* if it consists of clauses containing exactly one unnegated variable. We consider definite Horn formulas with clauses of size 3, i.e., with clauses of the form $(\bar{a} \vee \bar{b} \vee c)$, which can also be written as $a, b \rightarrow c$. Here a and b form the *body* of the clause and c is the *head* of the clause. Implication between a definite Horn formula φ and a definite Horn clause C can be decided by *forward chaining*: mark variables in the body of C and, while there is a clause in φ with all its body variables marked, mark its head variable as well. Then C is implied by φ iff its head gets marked.

We consider random definite Horn formulas with clauses of size 3 over n variables, where every clause is included with probability p. It follows directly from the results of [LMST09] that $p = (2 \ln n)/n$ is a threshold probability for the following property: *every pair of variables implies every other variable* (see also [DBC01] for a related result).

In this paper we consider the property that *some pair of variables implies every other variable*. This property is closely related to the property of propagation connectivity for 3-uniform undirected hypergraphs, introduced recently by Berke and Onsjö [BO09a]. They consider a marking process like forward

[†] This material is based upon work supported by the National Science Foundation under Grant No. CCF-0916708.

chaining, except that now a vertex can be marked if it is contained in an edge whose other two vertices are already marked. A 3-uniform undirected hypergraph is *propagation connected* if there is a pair of vertices such that the marking process, starting from that pair, marks every vertex. Berke and Onsjö showed that for $p < 1/(n(\log n)^2)$ random hypergraphs are *a.a.s.* not propagation connected [BO09a] and for $p > 1/(n(\log n)^{0.4})$ random hypergraphs are *a.a.s.* propagation connected [BO09b]. The first result proves a *lower bound* for the transition from random hypergraphs being *a.a.s.* not propagation connected to random hypergraphs being *a.a.s.* propagation connected, and the second result proves an *upper bound* for the transition. We use the terms lower and upper bound in a similar sense throughout the paper.

The Horn formula property mentioned above is equivalent to propagation connectivity for *directed* 3-uniform hypergraphs, where by a directed hypergraph we mean a hypergraph with each edge having a distinguished vertex called its head, and the other vertices called its body. (This is one of the possible definitions of a directed hypergraph. There are several other variants.) In the rest of the paper we use the terminology of propagation connectivity for directed hypergraphs instead of Horn formulas. We show that random directed 3-uniform hypergraphs for $p \le 1/(11n \ln n)$ are *a.a.s.* not propagation connected and for $p \ge (5 \ln \ln n)/(n \ln n)$ are *a.a.s.* propagation connected. The proofs are based on two versions of the "fanning-out process" (see, e.g., [Kar90, JŁR00]). For the upper bound we start the process by exploring a subset of the vertices and finding a maximal degree pair within that subset.

For the undirected hypergraph version of the problem, Coja-Oghlan, Onsjö and Watanabe concurrently and independently proved lower and upper bounds, both of the order $1/(n \ln n)$ [COOW10]. It appears that their argument can be adapted to yield order $1/(n \ln n)$ lower and upper bounds in the directed case as well. The proofs we present here are simpler.

The lower and upper bounds are presented in Section 3 and 4. In the closing section we mention a few open problems.

2 Preliminaries

We consider 3-uniform directed hypergraphs H with directed edges of the form $u, v \to w$. The pair (u, v) is the *body* of the edge and w is the *head* of the edge. Note that the body is an unordered pair. The *degree* of a pair (u, v) is the number of vertices w that form an edge $u, v \to w$ with the pair. We refer to vertex w as a *successor* of (u, v). The (u, v)-propagation connected component (or simply (u, v)-component) of H is the set of vertices marked by the marking process starting with (u, v).

The probability model where a random directed hypergraph is formed over the vertex set $[n] = \{1, \ldots, n\}$ by including each edge $u, v \to w$ independently with probability p is denoted by $\mathcal{DH}(n, p)$. For any monotone increasing property of directed hypergraphs the probability that the property holds for a random directed hypergraph drawn from $\mathcal{DH}(n, p)$ is a monotone non-decreasing function of p (see [Bol01, Th.2.1]).

We use the following versions of the Chernoff bounds [JŁR00].

Proposition 2.1 *If $X \in \mathrm{BIN}(n, p)$ and $t \ge 0$ then*

(a) $\Pr[X \ge \mathbb{E}[X] + t] \le \exp\left\{-\frac{t^2}{2\left(\mathbb{E}[X]+\frac{t}{3}\right)}\right\}$,

(b) $\Pr[X \le \mathbb{E}[X] - t] \le \exp\left\{-\frac{t^2}{2\mathbb{E}[X]}\right\}$.

3 A lower bound

We first give a lower bound for probabilities p such that a random directed hypergraph from $\mathcal{DH}(n,p)$ is *a.a.s.* propagation connected.

Theorem 3.1 *Let* $p \leq 1/(11n\ln n)$. *In a random directed hypergraph from* $\mathcal{DH}(n,p)$ a.a.s. *every propagation connected component has size at most* $11\ln n$.

Proof: By monotonicity we may assume $p = 1/(11n\ln n)$. The following process is used to explore $H \in \mathcal{DH}(n,p)$. Start with two sets $A_0 = \{u,v\}$ and $B_0 = \emptyset$. The sets A_i and B_i represent the sets of *discovered vertices* and *saturated pairs* at iteration i respectively, and put $m_i = |A_i|$. At iteration i of the process consider vertices u_i and v_i such that $u_i, v_i \in A_{i-1}$ and $(u_i,v_i) \notin B_{i-1}$. Find every edge $u_i, v_i \to w$ where $w \notin A_{i-1}$. Construct the set A_i so that it contains all vertices in set A_{i-1} plus all vertices w, where w is the head of an edge that was found in step i. Construct the set B_i by $B_i = B_{i-1} \cup \{(u_i,v_i)\}$. When every pair in A_i is saturated, we have discovered all the vertices in the component, and from then on we put $A_j = A_i$, $B_j = B_i$ for every $j > i$.

We need to show that this process stabilizes after a small number of steps with high probability. Define X_i to be the number of successors, in $V \setminus A_{i-1}$, of the pair (u_i,v_i) to be saturated. Each edge with body (u_i,v_i) and head in $V \setminus A_{i-1}$ is in the hypergraph with probability p, independently of the presence or absence of any other edge. Furthermore each such edge is considered at most once in the process. Thus $X_i \in \mathrm{BIN}(n - m_{i-1}, p)$.

Let $k = \lceil 11\ln n \rceil$. If the process generates at least k vertices then this must happen in the first $\binom{k-1}{2}$ iterations. Thus the probability of generating at least k vertices is at most

$$\Pr\left[\sum_{i=1}^{\binom{k-1}{2}} X_i \geq k-2\right]. \tag{1}$$

Let $X_i^+ \in \mathrm{BIN}(n,p)$ and replace the upper limit in the summation (1) by $\binom{k}{2}$ for convenience. Then, noting that $\sum_{i=1}^{\binom{k}{2}} X_i^+ \in \mathrm{BIN}\left(\binom{k}{2}n, p\right)$ and as such has mean $\binom{k}{2}np$, the probability (1) can be upper bounded by

$$\Pr\left[\sum_{i=1}^{\binom{k}{2}} X_i^+ \geq k-2\right] = \Pr\left[\sum_{i=1}^{\binom{k}{2}} X_i^+ \geq \binom{k}{2}np + k - 2 - \binom{k}{2}np\right].$$

Using the values of p and k we note that $np \sim 1/k$, giving $\binom{k}{2}np \sim k - 2 - \binom{k}{2}np \sim k/2$. Then the Chernoff bound (Proposition 2.1(a)) with $t = k - 2 - \binom{k}{2}np$ gives the upper bound $\exp\{-(3k)/16\} \sim \exp\{-(33\ln n)/16\} = o(n^{-2})$, which implies the theorem by the union bound. □

4 An upper bound

In this section we give a sufficient condition for probabilities p such that a random directed hypergraph from $\mathcal{DH}(n,p)$ is *a.a.s.* propagation connected.

Theorem 4.1 *For $p \geq (5 \ln \ln n)/(n \ln n)$ a random directed hypergraph from $\mathcal{DH}(n,p)$ is* a.a.s. *propagation connected.*

Proof: By monotonicity we may assume $p = (5 \ln \ln n)/(n \ln n)$. We use a modification of the process described above. First we consider *all* edges over the first $n/4$ vertices and find a highest-degree pair (u,v) in that subset. Starting from the successors of that pair we find a sufficiently large part of the rest of the component using a variant of the original process organized into phases as follows.

Let $m = \lceil (\ln n)/(\ln \ln n) \rceil$ and assume that we found a pair (u,v) with m successors $w_1, \ldots, w_m$ among the first $n/4$ vertices. Let $A_0 = \{w_1, \ldots, w_m\}$ be the initial set of discovered vertices and let C_0 be the $(3/4)n$ vertices not considered so far, forming the initial set of *available vertices*. In iteration i of the new process we pick an arbitrary set $D_{i-1} \subseteq C_{i-1}$ of $n/2$ available vertices, and we find all edges $u, v \to w$, where $u, v \in A_{i-1}$ and $w \in D_{i-1}$. If there are at least m distinct successors in D_{i-1} then let A_i be any m of these and put $C_i = C_{i-1} \setminus A_i$. Otherwise let $A_j = A_{i-1}$ for every $j \geq i$. We run this process for $\lceil \ln n \ln \ln n \rceil$ iterations.

The following lemma, analogous to bounds for graphs (see [Bol01, Ch.3]), gives a bound for the maximal degree of a pair in $H \in \mathcal{DH}(n,p)$. This lemma is stated for the smaller and simpler probability $1/(n \ln n)$, but applies also to larger p by monotonicity.

Lemma 4.2 *If $p = 1/(n \ln n)$, then the maximum degree of $H \in \mathcal{DH}(n,p)$ is* a.a.s. *at least* $(\ln 4n)/(\ln \ln 4n)$.

Proof: Let $d = \lceil (\ln 4n)/(\ln \ln 4n) \rceil$ and let the random variable Y_{ij} be the number of successors of pair (i,j) in H. Then $Y_{ij} \in \mathrm{BIN}\,(n-2, p)$ and since we are dealing with directed edges, the variables Y_{ij} are independent. Thus the probability that every degree is smaller than d is

$$(1 - \Pr[Y_{ij} \geq d])^{\binom{n}{2}} \leq \left(1 - \binom{n-2}{d} p^d (1-p)^{n-2-d}\right)^{\binom{n}{2}} < \left(1 - \frac{1}{2}\binom{n-2}{d} p^d\right)^{\binom{n}{2}},$$

if n is sufficiently large. For the last inequality we used the fact that $(1-p)^{n-2-d} = 1 - o(1)$. Using $1 - x < e^{-x}$, we need to show that

$$\left(\frac{pn}{d}\right)^d n^2 \to \infty.$$

This follows by taking logarithms and using the definitions of p and d. Specifically we use $(\ln 4n)/(\ln \ln 4n) \leq d \leq (2 \ln 4n)/(\ln \ln 4n)$ to get

$$d \ln\left(\frac{pn}{d}\right) + 2 \ln n > \frac{(\ln 4n)(\ln \ln \ln 4n - \ln 2)}{\ln \ln 4n} + (\ln n)\left(1 - \frac{\ln \ln n}{\ln \ln 4n}\right) - (\ln 4)\left(1 + \frac{\ln \ln n}{\ln \ln 4n}\right).$$

The expression on the right tends to infinity, since the first term tends to infinity, the second term is positive and the third term has a constant limit. □

We also use a version of a lemma of [BO09b] showing that *a.a.s.* every component is either small or contains every vertex. Such a statement holds for several probabilities p, but we state it here for $p = (5 \ln \ln n)/(n \ln n)$, as this property is not monotone. This lemma is similar to the gap theorem in [Kar90] and its proof is included for completeness.

Lemma 4.3 ([BO09b]) *If* $p = (5\ln\ln n)/(n\ln n)$ *then* a.a.s. *every propagation connected component has either size* n *or size less than* $(\ln n)^2$.

Proof: If a set of vertices is a propagation connected component then there can be no edges with body in the component and head outside. Thus the probability that there is a component of size k is at most

$$\binom{n}{k} \cdot (1-p)^{\binom{k}{2}(n-k)}.$$

We show that for $(\ln n)^2 \le k \le n-1$ this quantity is $o(1/n)$.
If $n/2 \le k \le n-1$ then replacing $\binom{n}{k}$ with $\binom{n}{n-k}$ gives the upper bound

$$\left(\frac{ne}{n-k}\right)^{n-k} \exp\left\{-p\binom{k}{2}(n-k)\right\} = \exp\left\{-(n-k)\left(p\binom{k}{2} - \ln\left(\frac{ne}{n-k}\right)\right)\right\}.$$

As $p\binom{k}{2} = \Omega\left(n\ln\ln n/\ln n\right)$, the probability is upper bounded by $\exp\{-\Omega\left(n\ln\ln n/\ln n\right)\}$.
Else $(\ln n)^2 \le k < n/2$ and the analogous calculation gives the upper bound

$$\exp\left\{-k\left(\ln k + \frac{p(k-1)(n-k)}{2} - (\ln n + 1)\right)\right\}.$$

Here $n-k$ can be replaced by $n/2$ and then substituting the values of p and k we can lower bound $\ln k + p(k-1)(n-k)/2 - (\ln n + 1)$ by $\Omega\left(\ln n\right)$. Since $k \ge (\ln n)^2$ we get an upper bound of the form $\exp\{-\Omega\left((\ln n)^3\right)\}$. □

Returning to the proof of Theorem 4.1, let us say that we are successful if we find a pair of degree m among the first $n/4$ vertices, we can run the iterative process for $\lceil \ln n \ln\ln n \rceil$ iterations, always finding m new vertices, and the event described in Lemma 4.3 occurs. In this case, after the last iteration we found a component of size $(\ln n)^2$, and by Lemma 4.3 the hypergraph is propagation connected.

The number Z_i of edges added in the ith iteration has distribution $\mathrm{BIN}\left(\binom{m}{2}\frac{n}{2}, p\right)$. Using the Chernoff bound (Proposition 2.1(b)) for the probability that there are fewer than m such edges we get

$$\Pr\left[Z_i < m\right] \le \exp\left\{-\frac{\left(\mathbb{E}[Z_i] - m\right)^2}{2\mathbb{E}[Z_i]}\right\} < \exp\left\{-\frac{\ln n}{41\ln\ln n}\right\}. \tag{2}$$

The last bound is due to $\mathbb{E}[Z_i] \sim m^2np/4 \sim (5\ln n)/(4\ln\ln n)$ and $\mathbb{E}[Z_i] - m \sim (\ln n)/(4\ln\ln n)$ which gives

$$\frac{\left(\mathbb{E}[Z_i] - m\right)^2}{2\mathbb{E}[Z_i]} \sim \frac{\ln n}{40\ln\ln n}.$$

Since we saturate more than one pair in an iteration it is possible that the same vertex is discovered by more than one edge. The probability of such a conflict is at most

$$\binom{\binom{m}{2}}{2} \cdot \frac{n}{2} \cdot p^2 = O\left(\frac{(\ln n)^2}{n(\ln\ln n)^2}\right). \tag{3}$$

If $Z_i \geq m$ and there are no conflicts in iteration i then we found at least m new vertices in that iteration. Hence, using Lemmas 4.2, 4.3 and the bounds (2) and (3), the probability of failure is

$$o(1) + O\left(\frac{(\ln n)^3}{n \ln \ln n} + \exp\left\{-\frac{\ln n}{41 \ln \ln n}\right\} \ln n \ln \ln n\right) = o(1).$$

□

The approach used in this paper seems to require edge probabilities of the order of magnitude given in Theorem 4.1, as the probability should be large enough to produce a sufficient number of newly discovered vertices, starting from an initial set of size determined by the maximum degree calculation of Lemma 4.2.

5 Open Problems

It follows from the result of [LMST09] mentioned in the introduction and Theorem 3.1 that the fraction of pairs (u, v) such that the (u, v)-component has size n grows from 0 to 1 over the interval between $p = \Omega\left(1/(n \ln n)\right)$ and $O\left(\ln n/n\right)$. It would be interesting to have more detailed information about growth over this interval.

Forward chaining can be studied for different ranges of the parameters. For example, [MIDV07] gives phase transition results on forward chaining where the marking process is started from a positive fraction of the vertices and p is of the form c/n^2. Forward chaining could be studied beyond just determining the size of the component produced. The process producing a propagation connected component can also be viewed, using the terminology of propositional logic, as a resolution derivation of consequences, or using the terminology of directed graphs, as a hyperpath [AIL$^+$10]. Considering combinatorial parameters of the structure of a propagation connected component, such as its depth, could also be of interest from the point of view of knowledge base applications.

The Internet and other large networks motivated the study of random graphs in models different from the standard models of fixed edge probabilities or fixed number of edges. For evolving knowledge bases, modeled by random Horn formulas, we are not aware of any such work. A possible choice would be to consider random subformulas of a given formula (corresponding to 'true' knowledge). For random graphs such a model has been studied, e.g., in [CH07].

6 Acknowledgements

The authors would like to thank the referee for a careful reading of this paper and useful suggestions.

References

[AIL$^+$10] Giorgio Ausiello, Giuseppe F. Italiano, Luigi Laura, Umberto Nanni, and Fabiano Sarracco. Classification and traversal algorithmic techniques for optimization problems on directed hyperpaths. Technical Report n.18, Dipartimento di Informatica e Sistemistica "Antonio Ruberti", Università di Roma "La Sapienza", 2010.

[BO09a] Robert Berke and Mikael Onsjö. Propagation connectivity of random hypergraphs. In *Stochastic Algorithms: Foundations and Applications, SAGA*, pages 117–126. Springer LNCS 5792, 2009. An update appears in [BO09b].

[BO09b] Robert Berke and Mikael Onsjö. Propagation connectivity of random hypergraphs. Technical Report 1342-2812, Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, October 2009.

[Bol01] Béla Bollobás. *Random Graphs*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2001.

[CH07] Fan R. K. Chung and Paul Horn. The spectral gap of a random subgraph of a graph. *Internet Mathematics*, 4(2):225–244, 2007.

[COOW10] Amin Coja-Oghlan, Mikael Onsjö, and Osamu Watanabe. Propagation connectivity of random hypergraphs. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, volume 6302 of *Lecture Notes in Computer Science*, pages 490–503. Springer Berlin / Heidelberg, 2010.

[DBC01] Paul E. Dunne and Trevor J. M. Bench-Capon. A sharp threshold for the phase transition of a restricted satisfiability problem for Horn clauses. *J. Log. Algebr. Program.*, 47(1):1–14, 2001.

[DV06] Demetrios D. Demopoulos and Moshe Y. Vardi. The phase transition in the random HornSAT problem. In *Computational Complexity and Statistical Physics (Santa Fe Institute Studies in the Sciences of Complexity Proceedings)*, pages 195–220. Oxford University Press, Inc., 2006.

[Ist02] Gabriel Istrate. The phase transition in random Horn satisfiability and its algorithmic implications. *Random Struct. Algorithms*, 20(4):483–506, 2002.

[JŁR00] Svante Janson, Tomasz Łuczak, and Andrzej Ruciński. *Random Graphs*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, 2000.

[Kar90] Richard M. Karp. The transitive closure of a random digraph. *Random Struct. Algorithms*, 1(1):73–94, 1990.

[LMST09] Marina Langlois, Dhruv Mubayi, Robert H. Sloan, and György Turán. Combinatorial problems for Horn clauses. In *Graph Theory, Computational Intelligence and Thought*, pages 54–65. Springer LNCS 5420, 2009.

[MIDV07] Cristopher Moore, Gabriel Istrate, Demetrios D. Demopoulos, and Moshe Y. Vardi. A continuous-discontinuous second-order transition in the satisfiability of random Horn-SAT formulas. *Random Struct. Algorithms*, 31(2):173–185, 2007.