Combinatorial problems for Horn clauses*

Marina Langlois
U. Illinois at Chicago
mirodo1@uic.edu

Dhruv Mubayi
U. Illinois at Chicago and
Carnegie Mellon University
mubayi@math.cmu.edu

Robert H. Sloan
U. Illinois at Chicago
sloan@uic.edu

György Turán
U. Illinois at Chicago and
Hungarian Academy of
Sciences & University of Szeged
gyt@uic.edu

Abstract

Given a family of Horn clauses, what is the minimal number of Horn clauses implying all other clauses in the family? What is the maximal number of Horn clauses from the family without having resolvents of a certain kind? We consider various problems of this type, and give some sharp bounds. We also consider the probability that a random family of a given size implies all other clauses in the family, and we prove the existence of a sharp threshold.

1 Introduction

Horn formulas form a basic framework for knowledge representation, being an expressive and tractable fragment of logic. They have been studied from many different aspects, such as reasoning and learning (Angluin, Frazier, & Pitt 1992; Kleine Büning & Lettmann 1999).

In our recent work (Langlois, Sloan, & Turán 2006; 2007; Langlois *et al.* 2007), we studied related problems on Horn formulas in the context of Horn approximation and belief revision. Motivated by applications such as the Open Mind Common Sense (Singh 2002) project for the acquisition of commonsense knowledge bases, we formulated the KnowBLe (Knowledge Base Learning) problem on synthesizing learning and belief revision. The objective is to learn a Horn formula in the model of learning with entailment, using a learning algorithm which updates its hypotheses in a rational manner in the spirit of the AGM paradigm (Alchourrón, Gärdenfors, & Makinson 1985).

In order to analyze various approaches to this problem, it would be useful to have a good understanding of the combinatorial and probabilistic properties of Horn formulas, such as how many additional clauses are implied and how many resolution steps can be formed. In this paper we consider some combinatorial problems of this sort.

Given a family of Horn clauses, what is the minimal number of Horn clauses needed to imply all Horn clauses in the family? What is the maximal number of Horn clauses in the family such that no resolution steps of a certain kind can be performed? These questions also appear to be of independent interest in combinatorics, as related questions about hypergraphs are much studied in extremal hypergraph theory.

As an interesting basic case, we discuss definite Horn clauses of size 3 in most of the paper (we briefly discuss the simple case of definite Horn clauses of size 2 as well, as it provides some interesting analogies). In the intended commonsense knowledge base application it seems reasonable to assume that the knowledge base contains non-unit definite clauses (in contrast to other applications where the knowledge base is used to derive facts from other facts).

It is noted that if a set of definite, size-3 Horn formulas implies a definite, size-3 Horn clause, then that clause has a resolution derivation using intermediate clauses of size at most 4 (and size-4 intermediate clauses may be necessary). The minimal number of definite, size-3 Horn clauses implying all definite Horn clauses of size 3 over n variables is determined exactly. At the other end, asymptotically sharp bounds are given for the maximal number of definite, size-3 Horn clauses over n variables, such that no resolution (resp., no resolution giving a resolvent of size 3, and no resolution giving a resolvent of size 4) can be performed among those clauses.

We also consider the probability that a given number of random definite, size-3 Horn clauses imply all other definite, size-3 Horn clauses. It is shown that this probability has a sharp threshold.

The paper is organized as follows. Section 2 gives some preliminaries, Section 3 discusses the case of definite Horn clauses of size 2 and Section 4 gives the bound on the size of intermediate clauses. Section 5 is on the minimal number of definite, size-3 Horn clauses implying all definite, size-3 Horn clauses. The bounds for the maximal number of definite, size-3 Horn clauses without different kinds of resolvents are contained in Section 6. Random formulas are considered in Section 7. We make a few final remarks in Section 8.

2 Preliminaries

We use standard terms from propositional logic such as literal and clause. Formulas are over n variables, and the variables are $X_n = \{x_1, \ldots, x_n\}$. A clause is *Horn* (resp., *definite Horn*) if it contains at most one (resp. exactly one) unnegated literal. We will generally write the Horn clause $(\bar{x} \vee \bar{y} \vee z)$ in the form $x, y \rightarrow z$. For a definite Horn clause

*This material is based upon work supported by the National Science Foundation under Grants No. CCF-0431059 and DMS-0653946.

C, let Body(C) be the set of variables corresponding to the negated literals in C and let Head(C) be the unnegated variable of C.

The *size* of a clause is the number of literals it contains. We use $\mathcal{D}_k^n$ to denote the collection of all size-k definite Horn clauses on n variables. Its size is

$$|\mathcal{D}_k^n| = k \cdot \binom{n}{k} \ , \tag{1}$$

which is $\Theta(n^k)$ for constant k.

A *(definite) Horn formula* is a conjunction—or a set, whichever view is more convenient—of (definite) Horn clauses.

A clause C is an *implicate* of a Boolean formula φ if every assignment satisfying φ also satisfies C; clause C is a *prime implicate* if it is an implicate but none of C's sub-clauses is an implicate.

We say that two clauses have an *opposing literal* when there is a variable that appears negated in one clause and unnegated in the other. A pair of Horn clauses can have either zero, one, or two opposing literals. We define the familiar operation of *resolution* for the case of Horn clauses to apply to a pair of Horn clauses that have exactly one opposing literal. Let C_1 and C_2 be such Horn clauses, and assume w.l.o.g. that C_1 is definite with its head being the opposing literal: Head$(C_1) \in$ Body(C_2). Resolution returns the clause $(\text{Body}(C_1) \cup \text{Body}(C_2) \setminus \{\text{Head}(C_1)\}) \to \text{Head}(C_2)$, which is called the *resolvent of C_1 and C_2*. Thus the resolvent of the two clauses $(a, b \to c)$ and $(c, d \to e)$ is $(a, b, d \to e)$. The resolvent of two definite, size-3 Horn clauses has size 3 or 4, referred to as a *3-resolvent*, resp., a *4-resolvent*. The resolvent of two definite size-2 Horn clauses always has size 2. A set of clauses F is called *resolvent-free* if no two clauses in F can be resolved.

We will use standard facts about Horn resolution, such as that every prime implicate of a Horn formula is a Horn clause, and we will also refer to the standard procedure of forward chaining, which can also be viewed as a unit resolution proof procedure (e.g., (Kleine Büning & Lettmann 1999; Russell & Norvig 2003)).

3 Definite Horn formulas with size-2 clauses

In this section we consider some extremal problems for definite Horn formulas with size-2 clauses. A definite, size-2 Horn clause $a \to b$ can be thought of as a directed edge (a, b), so definite Horn formulas with size-2 clauses can be viewed as directed graphs.

Proposition 1. *There is a subset of $\mathcal{D}_2^n$ of size n that has every clause in $\mathcal{D}_2^n$ as an implicate, and no smaller size subset has this property.*

Proof. The formula

$$(x_1 \to x_2) \wedge (x_2 \to x_3) \wedge \cdots \wedge (x_n \to x_1)$$

of clauses forming a cycle implies every size-2 definite Horn clause: For $i < j$, the clause $x_i \to x_j$ can be obtained by resolving $(x_i \to x_{i+1}), (x_{i+1} \to x_{i+2}), \ldots, (x_{j-1} \to x_j)$, two at a time in order. If instead $i > j$, then resolve $(x_i \to x_{i+1}), (x_{i+1} \to x_{i+2}), \ldots, (x_{n-1} \to x_n), (x_n \to x_1), (x_1 \to x_2), \ldots, (x_{j-1} \to x_j)$. At least n clauses are needed, as otherwise there is a variable that never appears as a head, and there is no way to obtain implicates having that head. □

Proposition 2. *If $F \subseteq \mathcal{D}_2^n$ is resolvent-free, then $|F| \leq \left\lfloor \frac{n^2}{4} \right\rfloor$ for $n \geq 3$, and the bound is sharp.*

Proof. Partition the set X_n of variables into sets A and B, and consider all clauses of the form $a \to b$ with $a \in A$ and $b \in B$. Clearly, this is a resolvent-free family. The number of clauses is maximized if $|A| = \lfloor \frac{n}{2} \rfloor$ and $|B| = \lceil \frac{n}{2} \rceil$, giving a family of the claimed size.

Now we show that the bound is the largest possible for $n \geq 3$. The cases $n = 3, 4$ are trivial. In digraph terms, we want the directed graph on n vertices with a maximum number of edges, having no simple path of length 2. If there is cycle of length two then its vertices cannot be incident to any other edge and the statement follows by induction. Otherwise, every vertex has either in-degree 0 or out-degree 0, and so the graph is a subgraph of a complete directed bipartite graph described above. □

4 Size of intermediate clauses in resolution

A resolution derivation of a short clause from a formula consisting of short clauses may contain large intermediate clauses. It is a basic observation with far-reaching implications that in some cases large intermediate clauses are unavoidable (Ben-Sasson & Wigderson 2001; Haken 1985). A trivial example for a class of clauses where such a phenomenon cannot occur is size-2 clauses, as every resolvent of such clauses has size at most 2. We note that there is also no blow-up of intermediate clauses for Horn formulas.

Theorem 3. *Let φ be a definite Horn formula with clauses of size at most 3. Then any size-3 prime implicate of φ has a resolution derivation where all clauses occurring in the proof have size at most 4.*

Proof sketch. Since φ is definite, all its resolvents and hence all its prime implicates must be definite Horn clauses. Assume that $C = a, b \to c$ is the implied clause. Then there is a forward chaining derivation of c from $\varphi \wedge a \wedge b$. In this derivation, each resolvent is shorter than its non-unit parent by one. Thus intermediate clauses all have size at most 2. Now omit any resolutions that used a or b. This new resolution derivation contains the same clauses as the original one, except that some clauses have a and/or b added to their body. Intermediate clause sizes could therefore be as large as 4. The final clause of this derivation, which is an implicate of φ, could in general have any subset of $\{a, b\}$ as its body and c as its head. However, since C is a prime implicate, the final clause must be C. □

The bound of 4 cannot be improved, as there may be size-3 prime implicates where we must use some intermediate clause of size 4. For example, we must use a size-4 intermediate resolvent to derive the prime implicate $a, b \to f$ of the

Horn formula

$$(a, c \to e) \wedge (b, d \to c) \wedge (d, e \to f) \wedge (a, b \to d).$$

Theorem 3 can be generalized, for example, to the following statement, using the same argument.

Corollary 4. *Let φ be a definite Horn formula with clauses of size at most s. Then any prime implicate of φ with t variables in its body has a resolution derivation where all intermediate clauses occurring in the proof have size at most $s - 1 + t$.*

5 Small formulas with all implicates

In this section we consider the problem of finding the smallest family of definite size-3 Horn clauses implying every clause in $\mathcal{D}_3^n$, and as in Proposition 1, we find the exact minimum.

Theorem 5. *There is a subset of $\mathcal{D}_3^n$ of size $\binom{n}{2}$ that has every clause in $\mathcal{D}_3^n$ as an implicate, and no smaller size subset has this property.*

Proof. To show that $\binom{n}{2}$ clauses are sufficient, we exhibit a set $S_n \subseteq \mathcal{D}_3^n$ of this size and demonstrate that S_n implies all definite size-3 clauses. Each clause of S_n is in one of three categories:

I. $x_i, x_j \to x_{i+1}$, for $i \leq n-2$ and $i+1 < j$,
II. $x_i, x_{i+1} \to x_{i+2}$, for $i \leq n-2$,
III. $x_{n-1}, x_n \to x_1$.

Note that S_n can be viewed as the size-3 analog of the directed cycle considered in Proposition 1.

All definite Horn clauses of size 3 are satisfied by the all 1's vector, the all 0's vector and all unit vectors. Call these vectors *standard*.

Assume for contradiction that $C \in \mathcal{D}_3^n$ is not implied by S_n. Then there is a truth assignment that satisfies S_n and falsifies C. This truth assignment has at least two 1's (corresponding to variables in Body(C)) and at least one 0 (namely, Head(C)), thus it is non-standard. Therefore, it is sufficient to show that every non-standard vector falsifies at least one clause of S_n. A non-standard vector can have the following forms (using regular expressions):

1. (0+1)*10(0+1)*1(0+1)*,
2. (0+1)*110(0+1)*,
3. 0(0+1)*11.

Vectors of form 1 falsify clauses of class I, vectors of form 2 falsify clauses of class II, and vectors of form 3 falsify clauses of class III.

For the lower bound, note that resolution of two definite clauses of size at least 3 does not produce clauses with any new bodies of size 2. Therefore a set of clauses implying all other clauses must contain all possible bodies. □

Incidentally, an examination of the resolutions needed shows that in order to derive all the clauses of $\mathcal{D}_3^n$ from S_n one does not need any size-4 intermediate clauses, unlike the general case given by Theorem 3.

Theorem 6. *Every clause $C \in \mathcal{D}_3^n$ can be derived from S_n with every intermediate clause being in $\mathcal{D}_3^n$.*

Proof sketch. A careful series of inductions shows that eventually all clauses can be derived. Table 1 gives a hint of the idea for each of the $\binom{n-1}{2}$ clauses with head x_h, for some arbitrary h. Starting at the leftmost column, with the boxed clause, we can derive all the clauses above the boxes, up through column $h+1$. Next these can be used to obtain the rest of the clauses in the leftmost columns (going "down" these same columns.) Two similar steps then handle the right columns. □

6 Large formulas without resolvents

A set of clauses $F \subseteq \mathcal{D}_3^n$ is *3-resolvent-free* (resp. *4-resolvent-free*) if no two of its clauses can be resolved to produce a 3-resolvent (resp., 4-resolvent). In this section we prove upper bounds on the size of resolvent-free, 3-resolvent-free, and 4-resolvent-free clause sets. First, we formulate a few technical lemmas.

6.1 Duplicates and Escher configurations

Definite, size-3 clauses C_1, C_2 form a *duplicate* if they contain the same set of variables. Thus duplicate clauses are of the form $a, b \to c$ and $a, c \to b$.

Lemma 7. *Let $F \subseteq \mathcal{D}_3^n$ be 3-resolvent-free. Then we can delete at most $\binom{n}{2}$ clauses from F such that no duplicates remain.*

Proof. It is sufficient to show that if clause $a, b \to c$ occurs in a duplicate in F, then F cannot contain another clause with the same body. Indeed, such a clause $a, b \to d$ gives a size-3 resolvent with both $a, c \to b$ and $b, c \to a$, and one of these clauses occurs in F. □

Lemma 8. *Let $F \subseteq \mathcal{D}_3^n$ be 4-resolvent-free. Then we can delete at most $(3/2)n^2$ clauses from F such that no duplicates remain.*

Proof. Consider the weighted undirected graph G on the vertex set X_n with an edge (b, c) for every pair of duplicates $a, b \to c$ and $a, c \to b$. The weight of such an edge (b, c) is the number of such vertices a. If we delete a clause from each such pair then no duplicates remain. Therefore, it is sufficient to prove the claimed upper bound for the sum of the edge weights in G.

We claim that the edges of G having weight at least 3 are independent. Assume that (b, c) and (c, d) both have weight at least 3. Then there is a vertex e such that $b, e \to c$ is in F, and there is a vertex f such that $c, f \to d$ is in F. For the last assertion we use the fact that (c, d) has weight at least 3, as f then can be chosen to be different from b and e. But the two clauses can be resolved to produce a 4-resolvent.

Hence the number of edges in G with weight at least 3 is at most $n/2$. Every weight is at most n, so the total weight of edges in G is at most $(n^2/2) + 2\binom{n}{2} \leq (3/2)n^2$. □

n	n-1	...	h+1	h-1	...	2
$x_1, x_n \to x_h$	$x_1, x_{n-1} \to x_h$	...	$x_1, x_{h+1} \to x_h$	$x_1, x_{h-1} \to x_h$	...	$x_1, x_2 \to x_h$
$x_2, x_n \to x_h$	$x_2, x_{n-1} \to x_h$	...	$x_2, x_{h+1} \to x_h$	$x_2, x_{h-1} \to x_h$	...	
$x_3, x_n \to x_h$	$x_3, x_{n-1} \to x_h$	...	$x_3, x_{h+1} \to x_h$	$x_3, x_{h-1} \to x_h$	...	
...	...	...	...	...	...	
$\boxed{x_{h-1}, x_n \to x_h}$	$\boxed{x_{h-1}, x_{n-1} \to x_h}$	...	$x_{h-2}, x_{h+1} \to x_h$	$\boxed{x_{h-2}, x_{h-1} \to x_h}$	...	
$x_{h+1}, x_n \to x_h$	$x_{h+1}, x_{n-1} \to x_h$	...	$\boxed{x_{h-1}, x_{h+1} \to x_h}$		...	
...	...					
$x_{n-2}, x_n \to x_h$	$x_{n-2}, x_{n-1} \to x_h$	...			...	
$x_{n-1}, x_n \to x_h$					...	

Table 1: All definite size-3 Horn clauses with head x_h

Definite, size-3 clauses C_1, C_2 form an *Escher configuration* if Head$(C_1) \in$ Body(C_2) and Head$(C_2) \in$ Body(C_1). (The name is inspired by Escher's *Drawing Hands* , though here it is heads rather than hands that are on one another.) Note that such a pair of clauses cannot be resolved.

Lemma 9. *Let $F \subseteq \mathcal{D}_3^n$ be resolvent-free. Then we can delete at most $2n^2$ clauses from F such that no Escher configurations remain.*

Proof. Consider the undirected graph G on the vertex set X_n with an edge $(\text{Head}(C_1), \text{Head}(C_2))$ for every pair of clauses $C_1, C_2 \in F$ forming an Escher configuration. We claim that every vertex of this graph has degree at most 2. Assume that d has neighbors a, b, c in G. Then F contains clauses

$$(a. \to d), (d. \to a), (b. \to d), (d. \to b), (c. \to d), (d. \to c),$$

where the dots correspond to one additional literal in each body. One can use a "sudoku" argument to derive a contradiction. If the second and third clauses cannot be resolved then the third clause must be $b, a \to d$. Similarly, if the fourth and fifth (resp., sixth and first) clauses cannot be resolved then the fifth (resp., first) clause must be $c, b \to d$ (resp., $a, c \to d$). But then the first and fourth clauses can be resolved.

Thus G has at most n edges. Every edge corresponds to at most $2(n-2)$ clauses (those obtained by adding a second literal to the bodies), and so the bound of the lemma follows. □

6.2 No resolvents

Let us partition the set of variables X_n into set A and B, and consider the set of clauses of the form $a, b \to c$ with $a, b \in A$, $c \in B$. Clearly, this is a resolvent-free family of definite, size-3 Horn clauses, which can be viewed as the size-3 analog of the complete directed bipartite graph of Section 3. The number of clauses in the family is $\binom{m}{2}(n-m)$, where $|A| = m$. This quantity is maximized for m with $|m - 2n/3| \leq 1$, and the maximum is

$$p(n) = \frac{4}{9}\binom{n}{3} + O(n^2).$$

The family constructed for the optimal value of m thus has size $p(n)$. We now show that this size is asymptotically optimal.

Theorem 10. *If $F \subseteq \mathcal{D}_3^n$ is resolvent-free then $|F| \leq p(n) + O(n^2)$.*

Proof. Let $F \subseteq \mathcal{D}_3^n$ be resolvent-free. Applying Lemma 7, we can delete $O(n^2)$ clauses such that no Escher configuration remains. In the remaining set F' of clauses, no variable can occur in a body of a clause and in the head of another clause, as those two clauses would either be resolvable or form an Escher configuration. Thus every variable is either head only, or body only, or neither. Thus F' is a subfamily of some family obtained by the above construction, and so its size is at most $p(n)$. □

6.3 No resolvents of size 3

Let us again partition the set of variables X_n into sets A and B, and this time consider the set of clauses of the form $a, b \to c$ with $a, b \in A$, $c \in B$, or $a, b \in B$, $c \in A$. This is a 3-resolvent-free family of definite, size-3 Horn clauses. (On the other hand, there are many resolvents of size 4.) The number of clauses in the family is $\binom{m}{2}(n-m) + \binom{n-m}{2}m$, where $|A| = m$. This quantity is maximized for m with $m = \lfloor n/2 \rfloor$, and the maximum is

$$q(n) = \frac{3}{4}\binom{n}{3} + O(n^2).$$

The family constructed for the optimal value of m thus has size $q(n)$. We now show that this size is asymptotically optimal.

Theorem 11. *If $F \subseteq \mathcal{D}_3^n$ is 3-resolvent-free then $|F| \leq q(n) + O(n^2)$.*

Proof. Let $F \subseteq \mathcal{D}_3^n$ be 3-resolvent-free. Applying Lemma 9, we can delete $O(n^2)$ clauses such that no duplicates remain. Let $a, b \to c$ be a clause in the remaining family F'. Then no clause in F' can have body $a, c \to$ or $b, c \to$, as any such clause would either produce a 3-resolvent with $a, b \to c$, or form a duplicate with it.

Consider the undirected graph G with vertices X_n and edges corresponding to the bodies of clauses in F', and let t be the number of vertex triples containing precisely one edge of G. The remark above implies that $|F'| \leq t$.

Therefore we get the required upper bound on $|F|$ by noting that the number of triples containing precisely one edge

of G is at most

$$\frac{1}{2}\sum d(v)(n-1-d(v)) \leq \frac{n}{2}\left(\frac{n-1}{2}\right)^2.$$

□

6.4 No resolvents of size 4

The families constructed in Section 6.2 have no resolvents, thus the same construction trivially shows that there are families of size $p(n)$ without 4-resolvents. Our final result shows that this is asymptotically optimal even for 4-resolvent-free families. This theorem thus supersedes Theorem 10; on the other hand, its proof is considerably more complicated and uses difficult recent results from extremal hypergraph theory. A *3-uniform hypergraph* is specified by a set of vertices and a set of 3-element subsets of the set of vertices.

Theorem 12. *For every $\epsilon > 0$ and sufficiently large n, if $F \subseteq \mathcal{D}_3^n$ is 4-resolvent-free then $|F| \leq p(n) + \epsilon n^3$.*

Proof. Let $F \subseteq \mathcal{D}_3^n$ be a 4-resolvent-free set of clauses. Applying Lemma 8, we can delete $O(n^2)$ clauses from F such that no Escher configurations remain. Let the remaining set of clauses be F', and consider the 3-uniform hypergraph H obtained from F' by replacing every clause $a, b \to c$ with the triple $\{a, b, c\}$. From now on we omit curly braces for triples and write abc for simplicity.

Let T be the 3-uniform hypergraph $\{abc, abd, abe, cde\}$ and T' be the 3-uniform hypergraph obtained from T by duplicating vertices a and b. Thus T' consists of the 13 triples $abc, ab'c, a'bc, a'b'c, abd, ab'd, a'bd, a'b'd, abe, ab'e, a'be, a'b'e, cde$. By an *orientation* of this family we mean a family of 13 definite, size-3 Horn clauses, each containing the 3 variables of a different triple from T'.

Lemma 13. *Any orientation of T' contains two clauses with a resolvent of size 4.*

Proof. Consider an orientation of T'. We may assume by symmetry that cde is oriented as $c, d \to e$. Then the clauses $a, b \to e$ and $a', b' \to e$ must be present, otherwise we get a 4-resolvent with $c, d \to e$. Now considering the triple $ab'c$, we find that every orientation leads to a 4-resolvent. □

It follows from Lemma 13 that H contains no copy of T'. From this, in the next lemma, we conclude that H contains only "few" copies of T.

Lemma 14. *For sufficiently large n, every 3-uniform, n-vertex, T'-free hypergraph has at most $n^{4.5}$ copies of T.*

Proof. Assume that G has more than $n^{4.5}$ copies of T. For every triple cde, let us consider the set of pairs ab which form a copy of T in G. Triples that have fewer than $n^{3/2}$ such pairs contribute at most $\binom{n}{3}n^{3/2} < n^{4.5}$ copies of T. Thus there is a triple cde with at least $n^{3/2}$ such pairs. The pairs corresponding to such a triple form a cycle $aba'b'$ of length 4 (Kővári, Sós, & Turán 1954). But then $\{a, b, a', b', c, d, e\}$ forms a T' in G. □

Now we can apply a special case of a deep result, the hypergraph removal lemma (Gowers 2006; Nagle, Rödl, & Schacht 2006; Tao 2006) to show that one can delete a "few" edges from H such that *no* copies of T remain.

Lemma 15 ((Gowers 2006; Nagle, Rödl, & Schacht 2006; Tao 2006))**.** *For every $\epsilon > 0$ and sufficiently large n, if H is a 3-uniform, n-vertex hypergraph containing at most $n^{4.5}$ copies of T, then one can delete ϵn^3 edges of H such that no copies of T remain.*

The maximal number of edges in a 3-uniform hypergraph without a copy of T has been determined exactly by (Füredi, Pikhurko, & Simonovits 2005).

Lemma 16 ((Füredi, Pikhurko, & Simonovits 2005))**.** *For sufficiently large n, every 3-uniform, n-vertex, T-free hypergraph has at most $p(n)$ edges.*

Now, putting things together, we get that the original set of clauses F contains at most $p(n) + O(n^2) + \epsilon n^3$ clauses, proving the theorem. □

7 Random formulas

In this section we consider a probabilistic version of the problem studied in Section 5. Let $p(n, s)$ be the probability that the conjunction of s random clauses from $\mathcal{D}_3^n$ implies every clause from $\mathcal{D}_3^n$. (Each clause is drawn from the uniform distribution over $\mathcal{D}_3^n$.) Informally, the property of implying every clause has a *sharp threshold* if around a certain number of clauses its probability jumps from low to high over a short interval (see, e.g., (Friedgut 1999)). The following result shows that $n^2 \ln n$ is a sharp threshold for this property. Note that for definite, size-2 Horn clauses the analogous property is the strong connectivity of random digraphs, which has been studied for a long time (e.g., (Karp 1990; Palásti 1966; Uno & Ibaraki 1998)).

Theorem 17. *For every $\epsilon > 0$ there exists a $c > 0$ such that if n is sufficiently large then*
a) $p(n, n^2 \ln n - cn^2) < \epsilon$,
b) $p(n, n^2 \ln n + cn^2) > 1 - \epsilon$.

Proof sketch. We use the following facts about the coupon collector problem: the expected number of trials needed to collect all of m coupons is $m \ln m + \Theta(m)$, and its variance is $\Theta(m^2)$. (See, e.g., (Motwani & Raghavan 1995).) Part *a)* follows directly from these facts, the Chebyshev inequality and the observation used in Theorem 5 that having all $\binom{n}{2}$ possible bodies in the formula is a necessary condition for generating every clause in $\mathcal{D}_3^n$.

In order to prove part *b)* we use another observation of Theorem 5: in order to show that a random formula of a given size implies every clause with high probability, it is sufficient to show that with high probability it is falsified by every non-standard vector.

Let $\mathbf{F}_s$ be the conjunction of s random clauses from $\mathcal{D}_3^n$. For $2 \leq k \leq n-1$ let

$$q(n, k, s) = Pr(\text{some weight } k \text{ vector satisfies } \mathbf{F}_s),$$

where the weight of a vector is the number of its ones. We would like to prove upper bounds for $q(n, k, s)$.

A vector of weight k falsifies $\binom{k}{2}(n-k)$ clauses in $\mathcal{D}_3^n$, as the body of such a clause must contain variables set to 1, and the head of such a clause must be a variable set to 0. So

$$q(n,k,s) \leq \left(1 - \frac{\binom{k}{2}\cdot(n-k)}{3\cdot\binom{n}{3}}\right)^s \cdot \binom{n}{k}.$$

For $k=2$, a direct computation shows that for $s = n^2\ln n + \left(\frac{1}{2}\ln\frac{2}{\epsilon}\right)n^2$ it holds that

$$q(n,2,s) < e^{-\frac{2s}{n(n-1)}}\binom{n}{2} < \frac{\epsilon}{2}.$$

If $3 \leq k \leq n-2$ then for $s = n^2\ln n$ it holds that

$$\begin{aligned} q(n,k,s) &< e^{-\frac{\binom{k}{2}\cdot(n-k)}{3\cdot\binom{n}{3}}\cdot s + k\ln\left(\frac{e\cdot n}{k}\right)} \\ &< e^{k\left(-\frac{(k-1)(n-k)}{n}\cdot\ln n + 1 + \ln n\right)} < n^{-2}. \end{aligned}$$

If $k = n-1$ then

$$q(n,n-1,s) \leq \left(1 - \frac{1}{n}\right)^s \cdot n$$

and so for $s = n^2\ln n$ again it holds that $q(n,n-1,s) = o(1)$. Thus

$$\sum_{k=3}^{n-1} q(n,k,n^2\ln n) = o(1),$$

hence part *b)* of the theorem follows. □

8 Further comments

The proof of Theorem 12 can be strengthened to show that $p(n)$ is actually the exact maximum (and thus Theorem 10 is also sharp). This result will be contained in a future paper. Along the same lines, it would be interesting to show that $q(n)$ is the exact maximum in Theorem 11.

There are many open problems related to the ones discussed here. Extending the results to definite Horn clauses of size greater than 3 seems to be interesting. For size-3 clauses the problems could be reduced to questions about graphs in several cases. For larger sizes this may not be the case anymore. Instead, one may get questions about hypergraphs, which tend to be more difficult.

From the point of view of the intended knowledge base learning application it would be interesting to extend Theorem 17 in several different ways. What is the expected number of clauses implied by a random family of s clauses for s below $n^2\ln n$? Other questions involve the length of resolution proofs of implied clauses. For s in the range $n^2\ln n$ or higher, a random family of s clauses implies every other clause with high probability. What is the expected length of the shortest resolution derivation of clauses?

References

Alchourrón, C. E.; Gärdenfors, P.; and Makinson, D. 1985. On the logic of theory change: partial meet functions for contraction and revision. *J. Symbolic Logic* 50:510–530.

Angluin, D.; Frazier, M.; and Pitt, L. 1992. Learning conjunctions of Horn clauses. *Machine Learning* 9:147–164.

Ben-Sasson, E., and Wigderson, A. 2001. Short proofs are narrow–resolution made simple. *J. ACM* 48(2):149–169.

Friedgut, E. 1999. Necessary and sufficient conditions for sharp threshold of graph properties and the k-SAT problem. *J. Amer. Math. Soc.* 12:1017–1054.

Füredi, Z.; Pikhurko, O.; and Simonovits, M. 2005. On triple systems with independent neighbourhoods. *Combinatorics, Probability and Computing* 14:795–813.

Gowers, T. 2006. Quasirandomness, counting and regularity for 3-uniform hypergraphs. *Combin. Probab. Comput.* 15:143–184.

Haken, A. 1985. The intractability of resolution. *Theoret. Comput. Sci.* 39:297–308.

Karp, R. M. 1990. The transitive closure of a random digraph. *Random Structures and Algorithms* 1:73–94.

Kővári, T.; Sós, V. T.; and Turán, P. 1954. On a problem of K. Zarankiewicz. *Colloquium Math.* 3:50–57.

Kleine Büning, H., and Lettmann, T. 1999. *Propositional Logic: Deduction and Algorithms*. Cambridge University Press.

Langlois, M.; Sloan, R. H.; Szörényi, B.; and Turán, G. 2007. Horn formulas, decomposability and belief revision. Submitted for publication.

Langlois, M.; Sloan, R. H.; and Turán, G. 2006. Horn upper bounds of random 3-CNF: A computational study. In *Ninth Int. Symp. Artificial Intelligence and Mathematics*. Available on-line from URL `http://anytime.cs.umass.edu/aimath06/`.

Langlois, M.; Sloan, R. H.; and Turán, G. 2007. Horn upper bounds and renaming. In *Proc. SAT 2007: Tenth Int. Conf. Theory and Applications of Satisfiability Testing*, volume 4501 of *Lecture Notes in Computer Science*, 80–93.

Motwani, R., and Raghavan, P. 1995. *Randomized Algorithms*. Cambridge Univ. Press.

Nagle, B.; Rödl, V.; and Schacht, M. 2006. The counting lemma for regular k-uniform hypergraphs. *Random Structures and Algorithms* 28:113–179.

Palásti, I. 1966. On the strong connectedness of directed random graphs. *Studia Sci. Math. Hungar.* 1:205–214.

Russell, S., and Norvig, P. 2003. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition edition.

Singh, P. 2002. The public acquisition of commonsense knowledge. In *Proc. AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*.

Tao, T. 2006. A variant of the hypergraph removal lemma. *J. Combin. Theory Ser. A* 113:1257–1280.

Uno, Y., and Ibaraki, T. 1998. Reachability problems of random digraphs. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* E81-A:2694–2702.