# The Number of k-SAT Functions

Béla Bollobás[*]
Department of Mathematical Sciences
University of Memphis
Memphis TN 38152-6429, U.S.A.
and
Trinity College
Cambridge CB2 1TQ
U.K.

Graham R. Brightwell [†]
Department of Mathematics
London School of Economics
Houghton St.
London WC2A 2AE
U.K.

21 December 2001

CDAM Research Report LSE-CDAM-2001-11

**Abstract**

We study the number $\mathrm{SAT}(k;n)$ of Boolean functions of $n$ variables that can be expressed by a $k$-SAT formula. Equivalently, we study the number of subsets of the $n$-cube $\mathbf{2}^n$ that can be represented as the union of $(n-k)$-subcubes. In [3] the authors and Imre Leader studied $\mathrm{SAT}(k;n)$ for $k \leq n/2$, with emphasis on the case $k = 2$. Here, we prove bounds on $\mathrm{SAT}(k;n)$ for $k \geq n/2$; we see a variety of different types of behaviour.

[*] email: bollobas@msci.memphis.edu, partially supported by NSF grant DMS-9971788 and DARPA grant F33615-01-C-1900

[†] email: g.r.brightwell@lse.ac.uk

# 1 Introduction

Let $\{x_1, \dots, x_n\}$ be a collection of $n$ Boolean *variables*. With each variable $x$ is associated a pair of *literals* $x$ and $\overline{x}$. The literal $\overline{x}$ is True/False if the variable $x$ is False/True. A *$k$-clause* is a set of $k$ literals, and a *$k$-SAT formula* is a set $\{C_1, \dots, C_t\}$ of $k$-clauses. A *satisfying assignment* for a $k$-SAT formula is an assignment of True/False to each variable such that each clause $C_i$ contains at least one True literal. A $k$-SAT formula gives rise to a Boolean function of $(x_1, \dots, x_n)$, where the output is 1 if and only if the input is a satisfying assignment. A *$k$-SAT function* is a function that has such a representation by a $k$-SAT formula. Observe that any function that can be represented by clauses of length *at most $k$* is a $k$-SAT function.

The families of $k$-SAT functions are probably the most studied special classes of Boolean functions, and our aim is to investigate how rich these families are. In other words, we want to estimate the number $\mathrm{SAT}(k;n)$ of $k$-SAT functions of $n$ variables. In this paper, we are mostly interested in the range $k \geq n/2$; as we shall see, there is a sharp divide in the behaviour of the function SAT around $k = n/2$.

The number of $k$-SAT formulae is $2^{2^k\binom{n}{k}}$, for any $n$ and $k$, since the number of possible $k$-SAT clauses is $2^k\binom{n}{k}$. For $k \geq 0.2278n$ this is greater than $2^{2^n}$, the number of Boolean functions, which leaves open the possibility that most Boolean functions are $k$-SAT functions. In fact, it is quite easy to see that this is too much to expect, but it seems natural to investigate how far from the truth this is. (Notice that the phase transition in the behaviour of $\mathrm{SAT}(k;n)$ around $k = n/2$ is not revealed by this naive comparison of upper bounds.)

Another way to view any satisfiability problem is in terms of subsets of the discrete cube $\mathbf{2}^n$. Here, we are interested in the number of subsets of the cube that can be represented as unions of *$d$-subcubes*, i.e., sets of the form

$$\{\mathbf{x} \in \mathbf{2}^n : x_i = a_i \text{ for all } i \in I\},$$

where $I$ is a subset of $[n]$ of size $n-d$, and each $a_i$ is 0 or 1. To be precise, let $\mathrm{CU}(d;n)$ be the number of subsets of $\mathbf{2}^n$ that are unions of subcubes of dimension at least $d$. (Although we shall normally assume $d$ is an integer, $\mathrm{CU}(d;n)$ is defined for any real $d$.)

For $d$ integer, it is well-known that $\mathrm{CU}(d;n) = \mathrm{SAT}(n-d;n)$; to see this, identify the variables $x_1, \dots, x_n$ with the co-ordinates of the cube $\mathbf{2}^n$ – a set of assignments is a subset of the cube, and to say that a set is the complement of the set of satisfying assignments for a $k$-SAT formula is exactly to express it as a union of $(n-k)$-subcubes. Therefore each $k$-SAT function corresponds to a subset of $\mathbf{2}^n$ that can be represented as a union of $d$-subcubes, where $d = n-k$, and $\mathrm{CU}(d;n) = \mathrm{SAT}(n-d;n)$, as claimed.

Estimating $\mathrm{CU}(d;n)$ seems to us to be an interesting question from a combinatorial point of view. For instance, $2^{-2^n}\mathrm{CU}(2;n)$ is the probability that a random subset of the $n$-cube can be written as a union of 2-cubes. It is not too hard to see that this probability tends to zero as $n$ tends to infinity; in fact, we shall prove that it is about $\exp\left(-2^{\frac{1}{2}\lg^2 n}\right)$ – here and throughout, lg denotes the binary logarithm. In the language of satisfiability, this result states that only this very small fraction of Boolean functions of $n$ variables can be represented by an $(n-2)$-SAT formula. (As we shall also see, about 61% of the Boolean functions of $n$ variables can be represented by an $(n-1)$-SAT formula.)

Bollobás, Brightwell and Leader [3] studied the function $\mathrm{SAT}(k;n)$ for $k \le n/2$. The main purpose of the research in [3] was a detailed investigation of the number of 2-SAT functions (alternatively, sets that are unions of $(n-2)$-subcubes). Bollobás, Brightwell and Leader showed that $\mathrm{SAT}(2;n) = 2^{(1+o(1))n^2/2}$, thus answering a question raised independently by Ursula Martin. Note that the number of *monotone* 2-SAT formulae, i.e., formulae in which only positive literals appear, is exactly $2^{\binom{n}{2}}$, and these all correspond to different functions, so the difficulty lies in showing a matching upper bound on $\mathrm{SAT}(2;n)$. It is conjectured in [3] that almost all 2-SAT functions arise from monotone formulae, possibly after relabelling literals as positive/negative.

It is also conjectured in [3] that, for each fixed $k$, $\mathrm{SAT}(k;n) = 2^{(1+o(1))\binom{n}{k}}$. Again, the family of functions arising from monotone formulae gives a lower bound of this form.

For larger values of $k$, Bollobás, Brightwell and Leader proved the following "monotonicity" result:

$$\mathrm{SAT}(k;n)^{1/\binom{n}{k}} \le \mathrm{SAT}(k;m)^{1/\binom{m}{k}}, \quad \text{for } k \le m \le n,$$

Applying this with $m = 2k$, using only the trivial bound $\mathrm{SAT}(k;2k) \le 2^{2^{2k}}$, gives

$$2^{\binom{n}{k}} \le \mathrm{SAT}(k;n) = \mathrm{CU}(n-k;n) \le 2^{\sqrt{\pi(k+1)}\binom{n}{k}},$$

whenever $k \le n/2$. See [3] for details.

In the context of the results for $k = 2$ in [3], the above bound $\mathrm{SAT}(k;n) \le 2^{\sqrt{\pi(k+1)}\binom{n}{k}}$ may not seem especially impressive for fixed $k$, and it is surely only the first step towards proving the conjecture mentioned. It is probably more important that the inequality holds whenever $k \le n/2$ since, in combination with the results of this paper, it reveals a "phase transition" in $\mathrm{SAT}(k;n)$ as $k$ goes through $n/2$. Indeed, if $k = \alpha n$ with $\alpha < \frac{1}{2}$, then we have $\mathrm{SAT}(k;n) = 2^{o(2^n)}$; on the other hand we prove that, when $k = \alpha n$ with $\alpha > \frac{1}{2}$, there is a constant $\beta = \beta(\alpha) > 0$ such that $\mathrm{SAT}(k;n) \ge 2^{\beta 2^n}$. We also prove a lower bound on $\mathrm{SAT}(k;n)$ of the same general form.

In fact, we suspect that the "monotone formulae" lower bound might be roughly correct not just for fixed $k$, but whenever $k \le (\frac{1}{2} - \varepsilon)n$.

**Conjecture 1.1** *If $\varepsilon > 0$, and $k = k(n) \le (\frac{1}{2} - \varepsilon)n$, then $\mathrm{SAT}(k;n) = 2^{\binom{n}{k}(1+o(1))}$ as $n \to \infty$.*

For the rest of the paper we only consider the case where $k \ge n/2$. It is much more convenient in this context to work with the function $\mathrm{CU}(d;n) = \mathrm{SAT}(n-d;n)$, with $d \le n/2$.

Our results (and those from [3]) are summarised in the following theorem.

**Theorem 1.2**

(1) $d = 1$ : $\qquad \lg \mathrm{CU}(1;n) = 2^n - \frac{1}{2}\lg e + o(1).$

(2) $d \geq 2$ *constant*: $\qquad \lg \mathrm{CU}(d;n) = 2^n - 2^{\frac{d-1}{2}\lg^2 n + o(\lg^2 n)}.$

(3) $\log^{3/2} n \leq 2^{2^d/(d-1)} \leq n^{1/4}$ : $\qquad \lg \mathrm{CU}(d;n) = 2^n - \exp\left\{2^{2^d/(d-1)}(\log n)^{d/(d-1)} d 2^{O(1)}\right\}.$

(4) $6\sqrt{\lg n} \leq d = o(n)$ : $\qquad 2^n\left(1 - \frac{d\log(n/d)}{n}(1+o(1))\right) \leq \lg \mathrm{CU}(d;n) \leq 2^n\left(1 - \frac{d^3}{3n^3}\right).$

(5) $d = \alpha n, \alpha < 1/2$ : $\qquad 2^n\frac{1-2\alpha}{1-\alpha}\left(\frac{\alpha}{1-\alpha}\right)^{\alpha/(1-2\alpha)}(1+o(1)) \leq \lg \mathrm{CU}(d;n) \leq 2^n\left(1 - \frac{\alpha^3}{3}\right).$

(6) $n/2 \leq d \leq n-3$ : $\qquad \binom{n}{d} \leq \lg \mathrm{CU}(d;n) \leq \sqrt{\pi(n-d+1)}\binom{n}{d}.$

(7) $d = n-2$ : $\qquad \lg \mathrm{CU}(n-2;n) = \binom{n}{2}(1-o(1)).$

(8) $d = n-1$ : $\qquad \lg \mathrm{CU}(n-1;n) = \lg(3^n+1) \simeq n\lg 3.$

Of the assertions in Theorem 1.2, (1) (which is quite simple) is Theorem 2.2; (2) is Theorem 2.4 in the case $d = 2$ (where actually we prove a slightly stronger result) and Theorem 2.8 for $d \geq 3$. A slightly more precise version of (3) is Theorem 2.10; the range of $d$ covered here runs from about $\lg\lg\lg n$ to about $\lg\lg n$. The lower bounds in (4) and (5) are special cases of Theorem 3.1, while the upper bounds are contained in Theorem 4.2. We have already discussed (6) and (7), which are from [3], as is the easy final identity stated as (8).

In the next section, we study $\mathrm{CU}(d;n)$ where $d$ is constant or grows very slowly with $n$. In these cases, as set out in parts (1)–(3) of Theorem 1.2, we shall be able to estimate $\mathrm{CU}(d;n)$ fairly accurately.

Then we move on to larger values of $d$, establishing the lower bounds in Section 3 and the upper bounds in Section 4. At this point, we shall have completed the proof of Theorem 1.2.

In the last section we show that, except possibly for a countable set of real numbers $\alpha$, $2^{-n}\lg \mathrm{CU}(\alpha n;n)$ tends to a limit as $n \to \infty$.

We use a variety of techniques in the different sections. Of particular interest might be the use of entropy methods, as pioneered by Kahn [6, 7], in Section 4.

Although our results give a reasonable picture of the behaviour of $\mathrm{CU}(d;n)$, there are still plenty of gaps, especially in the transitions between the various ranges specified above. It would also be interesting to close the gaps between lower and upper bounds, especially in (4), (5) and (6) of Theorem 1.2.

# 2 Small dimensional cubes

In the case where $d$ is small, our approach is to take a random subset $S$ of $\mathbf{2}^n$, and ask for the probability that $S$ is a union of $d$-subcubes. If $x$ is any fixed point in $\mathbf{2}^n$, the probability that it is taken into $S$ but is not contained in a $d$-subcube in $S$ will turn out to be very small, although the expected number $\lambda$ of "bad" points $x$ will be large. We would hope that the probability that there are no bad points is approximately $e^{-\lambda}$. An especially versatile tool for proving results of this kind is Suen's Inequality [8] (or see Alon and Spencer [1], Theorem 8.7.1).

We need some terminology and notation. Let $(A_i) = \{A_i : i \in I\}$ be a family of events in an arbitrary probability space. A symmetric relation $\sim$ on $I$ defines a *superdependency graph* for the family $(A_i)$ if, whenever $J_1$ and $J_2$ are disjoint subsets of $I$ such that we never have $j_1 \sim j_2$ for $j_1 \in J_1$ and $j_2 \in J_2$, then any Boolean combination of the events $\{A_i : i \in J_1\}$ is independent of any Boolean combination of the events $\{A_i : i \in J_2\}$. Suen's Inequality is as follows.

**Theorem 2.1** *Suppose that $(A_i) = \{A_i : i \in I\}$ is a family of events in a probability space, and $\sim$ defines a superdependency graph for the family $(A_i)$. Let $Q = \prod_{i\in I}(1-\mathbb{P}(A_i))$. Then:* $B$

$$|\mathbb{P}(\textit{no } A_i \textit{ occurs}) - Q| \le Q\left(\exp\left(\sum_{i\sim j}\varphi(A_i, A_j)\right) - 1\right),$$

*where*

$$\varphi(A_i, A_j) = (\mathbb{P}(A_i \textit{ and } A_j) + \mathbb{P}(A_i)\mathbb{P}(A_j)) \prod_{\ell\sim i \textit{ or } \ell\sim j} \frac{1}{1-\mathbb{P}(A_\ell)}.$$

To illustrate the situation, let us describe the examples we are interested in. We shall take the underlying probability measure to be the uniform measure on subsets $S$ of $\mathbf{2}^n$. Let $I = \mathbf{2}^n$, set $x \sim_d y$ if $d(x,y) \le 2d$, and let $A_d(x)$ be the event that $x \in S$ but $x$ is not contained within a $d$-subcube in $S$. Noting that $A_d(x)$ depends only on whether or not those points of $\mathbf{2}^n$ within distance $d$ of $x$ are in $S$, we see that $\sim_d$ defines a superdependency graph for the family $(A_d(x)) = \{A_d(x) : x \in \mathbf{2}^n\}$.

Let us now begin our investigation of $\mathrm{CU}(d;n)$. If $d = 0$, then of course all the $2^{2^n}$ subsets of $\mathbf{2}^n$ can be represented as unions of $d$-subcubes, so $\mathrm{CU}(0;n) = 2^{2^n}$. If $d = 1$, we are asking how many sets can be represented as unions of 1-subcubes (edges). A set is a union of 1-subcubes if and only if it has no isolated points. So we are asking for the probability that a random subset $S$ of $\mathbf{2}^n$ contains no isolated points. This is straightforward to estimate.

**Theorem 2.2** *The probability that a random subset of $\mathbf{2}^n$ contains no isolated points is $e^{-1/2} + O(n^2 2^{-n}) \simeq 0.6065$. Therefore $\mathrm{CU}(1;n) = 2^{2^n}(e^{-1/2} + o(1))$ as $n \to \infty$.*

**Proof.** Let $S$ be a random subset of $\mathbf{2}^n$. We apply Suen's Inequality to the family $(A_1(x))$ and the superdependency graph given by $\sim_1$, defined as above. Note that $\mathbb{P}(A_1(x)) = 2^{-(n+1)}$, and so

$$Q = \prod_{x \in I}(1 - \mathbb{P}(A_x)) = \left(1 - 2^{-(n+1)}\right)^{2^n} = e^{-1/2} + O(2^{-n}).$$

If $d(x, y) = 1$, then $x$ and $y$ cannot both be isolated points of $S$, while if $d(x, y) = 2$ then the probability that $x$ and $y$ are both isolated points is exactly $2^{-2n}$. Hence if $x \sim_1 y$ then the quantity $\varphi(A_x, A_y)$ of Suen's Inequality is at most

$$\left(2^{-2n} + 2^{-2n-2}\right)\left(1 - 2^{-n-1}\right)^{-n^2} \leq 2^{-2n+1},$$

for sufficiently large $n$. Hence $\sum_{i \sim j} \varphi(A_i, A_j) \leq n^2 2^{-n}$, and therefore the probability that no $x$ is isolated is $e^{-1/2} + O(n^2 2^{-n})$, as claimed. □

The remaining results in this section are all proved in much the same way. Let $x$ be a point of the $r$-dimensional cube $\mathbf{2}^r$, and define a random subset $H$ by taking $x$ and all its neighbours, and each other point of $\mathbf{2}^r$ independently with probability 1/2. Then let $B_{r,d}$ be the event that $H$ does not contain a $d$-subcube including $x$. We see that

$$\mathbb{P}(A_d(x)) = \sum_{r=0}^{n} 2^{-(n+1)} \binom{n}{r} \mathbb{P}(B_{r,d}),$$

since the $r$-term is the probability that, in a random subset $S$ of $\mathbf{2}^n$, $x$ and exactly $r$ of its neighbours are in $S$, and $A_d(x)$ occurs.

We can also think of the random subset $H$ above as a random hypergraph $H$ on a set $R$ of $r$ vertices, where we take each subset of $R$ of size at least 2 independently with probability 1/2. A $d$-subcube then corresponds to a $d$-set and all of its subsets of size at least 2 being included in $H$.

From now on, we shall always be using Suen's Inequality in the same way, so it is convenient to state a lemma encapsulating what we obtain from the inequality, and in particular specifying a sufficient set of conditions for its use.

**Lemma 2.3** *Suppose that $d = d(n)$ is a function satisfying: $d = o(n)$, $\mathbb{P}(A_d(x)) = 2^{-n+o(n)}$, and $\mathbb{P}(B_{\lceil n^{3/4} \rceil, d}) \leq 2^{-2n}$ for sufficiently large $n$. Then*

$$\mathbb{P}(\textit{no } A_d(x) \textit{ occurs}) = \exp\left(-2^n \mathbb{P}(A_d(x)) + O(2^{-n+o(n)})\right).$$

**Proof.** We shall apply Suen's Inequality with the superdependency graph defined by $\sim_d$ on the family $(A_d(x))$. Note first that

$$Q = (1 - \mathbb{P}(A_d(x)))^{2^n} = \exp\left(-2^n \mathbb{P}(A_d(x)) + 2^{-n+o(n)}\right).$$

Next, for any $x$ and $y$,

$$\prod_{z \sim_d x \text{ or } z \sim_d y} \frac{1}{1 - \mathbb{P}(A_d(z))} = \left(\frac{1}{1 - 2^{-n+o(n)}}\right)^{O\left(\binom{n}{2d}\right)} = 1 + 2^{-n+o(n)},$$

since $\binom{n}{2d} = 2^{o(n)}$.

For the final ingredient, observe that, for any $x$ and $y$,

$$\begin{aligned}\mathbb{P}(A_d(x) \text{ and } A_d(y)) &\leq \mathbb{P}(|S \cap N(x)| < n^{3/4} \text{ and } |S \cap N(y)| < n^{3/4}) + \\ &\quad + \mathbb{P}(A_d(x) \mid |S \cap N(x)| \geq n^{3/4}).\end{aligned}$$

Note that $|N(x) \cup N(y)| \geq 2n - 2$, so the first term above is at most $2\binom{2n}{2\lfloor n^{3/4}\rfloor}2^{-2n+2} = 2^{-2n+o(n)}$, while the second is at most $\mathbb{P}(B_{\lceil n^{3/4}\rceil,d})$, which is at most $2^{-2n}$ by assumption. Hence $\mathbb{P}(A_d(x) \text{ and } A_d(y)) = 2^{-2n+o(n)}$.

These facts imply that $\varphi(A_d(x), A_d(y)) = 2^{-2n+o(n)}$ whenever $x \sim_d y$. Hence

$$\sum_{x \sim_d y} \varphi(A_d(x), A_d(y)) = 2^n \binom{n}{2d}(1 + o(1))2^{-2n+o(n)} = 2^{-n+o(n)}.$$

Suen's Inequality now tells us that

$$|\mathbb{P}(\text{no } A_d(x) \text{ occurs}) - Q| = O\left(Q2^{-n+o(n)}\right),$$

which implies the result. □

The conditions of Lemma 2.3 will usually be relatively easy to check; most of the work will be in estimating $\mathbb{P}(A_d(x))$. We can deal with the case $d = 2$ very quickly now.

**Theorem 2.4** *The probability that, in a random subset $S$ of $\mathbf{2}^n$, every point of $S$ lies in a 2-subcube in $S$ is*

$$\exp\left(-2^{\frac{1}{2}\lg^2 n - \lg n \lg\lg n + O(\lg n)}\right).$$

*Therefore*

$$\mathrm{CU}(2; n) = 2^{2^n} \exp\left(-2^{\frac{1}{2}\lg^2 n - \lg n \lg\lg n + O(\lg n)}\right).$$

**Proof.** Recall that $\mathbb{P}(A_2(x)) = 2^{-(n+1)} \sum_{r=0}^{n} \binom{n}{r}\mathbb{P}(B_r, 2)$. Since $\mathbb{P}(B_r, 2) = 2^{-\binom{r}{2}}$, we have

$$\mathbb{P}(A_2(x)) = 2^{-n+O(\lg n)} \max_r \left\{ \binom{n}{r} 2^{-\binom{r}{2}} \right\}.$$

This maximum is attained at $r = \lg n - \lg\lg n + O(1)$, and is equal to $2^{\frac{1}{2}\lg^2 n - \lg n \lg\lg n + O(\lg n)}$, so

$$\mathbb{P}(A_2(x)) = 2^{-n + \frac{1}{2}\lg^2 n - \lg n \lg\lg n + O(\lg n)}.$$

Also $\mathbb{P}(B_{\lceil n^{3/4}\rceil,2}) = 2^{-O(n^{3/2})} \leq 2^{-2n}$ for sufficiently large $n$.

Therefore we can apply Lemma 2.3 and obtain

$$\mathbb{P}(\text{no } A_d(x) \text{ occurs}) = \exp\left(-2^n \mathbb{P}(A_d(x))\right),$$

which is the required result. □

For higher values of $d$, the application of Lemma 2.3 will be equally straightforward, and the crucial task is to estimate the probability of the event $A_d(x)$, which we again approach via the events $B_{r,d}$.

Recall that, in a random hypergraph $H$ on an $r$-element set $R$, $B_{r,d}$ can be thought of as the event that there is no $d$-element set $D$ such that all subsets of $D$ of size at least 2 are in $H$. We can treat the subsets of size 2 in $H$ as the edges of a random graph $G_H$ on $R$, and note that a $d$-subcube induces a complete subgraph $K_d$ in $G_H$. Hence $\mathbb{P}(B_{r,d})$ is at least the probability that the random graph $G_H$ contains no $K_d$, which is at least $2^{-r^2/2(d-1)}$. (This is a lower bound on the probability that $G_H$ respects a particular partition into $d-1$ independent sets as equal in size as possible.)

We will get an upper bound on $\mathbb{P}(B_{r,d})$ using Szemerédi's Uniformity Lemma (see e.g., Bollobás [2]), of which we now remind the reader.

For a graph $G$, and disjoint subsets $Y$, $Z$ of $V(G)$, the *density* $\rho(Y,Z)$ is defined as the number of edges between $Y$ and $Z$ divided by $|Y|\,|Z|$.

Given a graph $G$ and $\varepsilon > 0$, an *$\varepsilon$-uniform pair* in $G$ is a pair of subsets $(Y,Z)$ of $V(G)$ such that, for any $S \subset Y$ and $T \subset Z$, with $|S| \geq \varepsilon|Y|$ and $|T| \geq \varepsilon|Z|$, we have

$$|\rho(S,T) - \rho(Y,Z)| < \varepsilon.$$

Now, given $G$ and $\varepsilon$, an *$\varepsilon$-uniform partition* of $G$ is a partition of $V(G)$ into sets $Y_0, Y_1, \ldots, Y_m$ with $|Y_0| \leq \varepsilon|V(G)|$, and $|Y_1| = \cdots = |Y_m|$, such that all but at most $\varepsilon m^2$ of the pairs $(Y_i, Y_j)$ with $1 \leq i < j \leq m$ are $\varepsilon$-uniform.

**Theorem 2.5** *For any $\varepsilon > 0$, and any integer $k$, there is some $K = K(k, \varepsilon)$ such that every graph $G$ on at least $k$ vertices has an $\varepsilon$-uniform partition into sets $Y_0, Y_1, \ldots, Y_m$ for some $m$ with $k \leq m \leq K$.*

We also use the following lemma, which is a version of a standard tool for use with Szemerédi's Uniformity Lemma. Its statement is designed for straightforward proof by induction on $d$, exactly as in, for instance, Theorem IV.6.30 of Bollobás [2].

**Lemma 2.6** *Take any constant $\delta$ with $0 < \delta \leq 1/2$. Let $G$ be a graph, and suppose that $Y_1, \ldots, Y_d$ are disjoint subsets of $V(G)$ of size at least $s$ such that, whenever $W_i$ is a subset of $V_i$ of size at least $\delta^d|Y_i|$, and $W_j$ is a subset of $V_j$ ($j \neq i$) of size at least $\delta^d|Y_j|$, we have $\rho(W_i, W_j) \geq \delta$. Then there are at least $(s/2)^d \delta^{\binom{d}{2}}$ copies of $K_d$ in $G$, each with one vertex in each $Y_i$.*

We now prove our promised upper bound on $\mathbb{P}(B_{r,d})$.

**Lemma 2.7** *For any integer $d \geq 3$, and any $\eta > 0$, there is some $r_0$ such that, for $r \geq r_0$,*

$$\mathbb{P}(B_{r,d}) \leq 2^{-r^2/2(d-1)+\eta r^2}.$$

**Proof.** We start by choosing parameters. Given $d$ and $\eta$, take $\delta \in (0, 1/10)$ such that $4\delta \log(\delta^{-1}) < \eta$. Set $\varepsilon = \delta^d$, $k = \lceil 1/\varepsilon \rceil$, and let $K = K(k, \varepsilon)$ be as in the statement of Theorem 2.5. Let $\gamma = ((1-\varepsilon)/2K)^d \, \delta^{\binom{d}{2}}$, and finally choose $r_0$ large enough to satisfy all of the following:

$$K^2 \leq \eta r_0^2/2; \quad 2(K+1) \leq 2^{\eta r_0/4}; \quad \left(1 - 2^{-2^d}\right)^{\gamma r_0 / \binom{d}{3}} \leq 2^{-1/2(d-1)}.$$

Now take any $r \geq r_0$, noting that $r$ also satisfies these inequalities, which we shall use as required without specific reference. As before, take a random hypergraph $H$ on a set $R$ of $r$ vertices, and let $G_H$ be the graph formed by the 2-sets in $H$. Let $\Gamma$ be the event that $G_H$ contains at least $\gamma r^d$ copies of $K_d$, and observe that

$$\mathbb{P}(B_{r,d}) \leq \mathbb{P}(\overline{\Gamma}) + \mathbb{P}(B_{r,d} \mid \Gamma).$$

Our plan is to bound both of these terms above. We start with $\mathbb{P}(\overline{\Gamma})$, so our aim is to show that the proportion of $r$-vertex graphs without $\gamma r^d$ copies of $K_d$ is suitably small.

Take an $\varepsilon$-uniform partition of $G_H$ into sets $Y_0, Y_1, \ldots, Y_m$, with $k \leq m \leq K$, as guaranteed by Theorem 2.5. Suppose that $|Y_1| = \cdots = |Y_m| = s$, and note that $s \leq r/m \leq \varepsilon r$.

For $1 \leq i < j \leq m$, call the pair $(Y_i, Y_j)$ *dense* if it is $\varepsilon$-uniform and $\rho(Y_i, Y_j) \geq 2\delta$. The set of dense pairs can be thought of as a graph $G'$ on $\{Y_1, \ldots, Y_m\}$.

Suppose that $G'$ contains a clique of size $d$, say on sets $Y_1, \ldots, Y_d$. Then $G_H$ satisfies the hypotheses of Lemma 2.6, so there are at least $(s/2)^d \delta^{\binom{d}{2}}$ copies of $K_d$ in $G_H$. Noting that $s \geq r(1-\varepsilon)/K$, and referring to the specification of $\gamma$, this means that $\Gamma$ occurs. Therefore $\mathbb{P}(\overline{\Gamma})$ is at most the proportion of graphs having an $\varepsilon$-uniform partition as specified without a clique of size $d$ in the graph $G'$ formed by the dense pairs.

There are at most $(K+1)^r 2^{K^2/2}$ ways of partitioning the vertex set $R$ and selecting a graph $G'$. If $G'$ does not contain a clique of size $d$, then – by Turán's Theorem – the total number of edges in $G'$ is at most $\frac{m^2}{2} \frac{d-2}{d-1}$. Given $G'$, the number of ways of constructing $G$ consistent with the prescription of dense pairs is at most

$$2^{|Y_0| r + m\binom{s}{2} + s^2\left(\frac{m^2}{2}\frac{d-2}{d-1} + \varepsilon m^2\right)} \binom{s^2}{2\delta s^2}^{m^2/2(d-1)}.$$

The terms here count, respectively: choices of edges incident with $Y_0$, choices of edges inside the parts of the partition, choices of edges across the dense pairs and those pairs that are not $\varepsilon$-uniform, and choices of (few) edges across the non-dense pairs. The quantity above is at most

$$2^{\frac{r^2}{2}(2\varepsilon + \varepsilon + \frac{d-2}{d-1} + 2\varepsilon)} \left(\frac{e}{2\delta}\right)^{\delta r^2/(d-1)} \leq 2^{\frac{r^2}{2}\frac{d-2}{d-1} + r^2\left(\frac{5\varepsilon}{2} + \frac{\delta}{2}\log(\delta^{-1}e/2)\right)} \leq 2^{\frac{r^2}{2}\frac{d-2}{d-1} + \delta \log(\delta^{-1}) r^2}.$$

So $\mathbb{P}(\overline{\Gamma})$ is at most

$$(K+1)^r 2^{K^2/2} 2^{\frac{r^2}{2}\frac{d-2}{d-1} + \delta\log(\delta^{-1})r^2} 2^{-\frac{r(r-1)}{2}} \leq 2^{-r} 2^{\eta r^2/4} 2^{\eta r^2/4} 2^{\frac{r^2}{2}\frac{d-2}{d-1} + \eta r^2/4} 2^{-r^2/2} 2^{r/2} \leq 2^{-r^2/2(d-1) + \frac{3}{4}\eta r^2}.$$

It remains to bound $\mathbb{P}(B_{r,d} \mid \Gamma)$. Suppose then that $\Gamma$ holds: i.e., there are at least $\gamma r^d$ copies of $K_d$ in $G_H$. Each 3-set is contained in at most $r^{d-3}$ of these copies of $K_d$, so each copy of $K_d$ shares a triangle with at most $\binom{d}{3} r^{d-3}$ other copies. Therefore, there is a collection of at least $\gamma r^3/\binom{d}{3}$ $d$-cliques in $G_H$, no two of which share a triangle. Let the vertex sets of these $d$-cliques be $W_1, \ldots, W_t$, and let $C_j$ be the event that every subset of $W_j$ of size at least 2 is in the random hypergraph $H$, for $j = 1, \ldots, t$. The events $C_j$ are independent, and each has probability at least $2^{-2^d}$, so the probability that none of them occurs is at most

$$\left(1 - 2^{-2^d}\right)^{\gamma r^3/\binom{d}{3}} \leq 2^{-r^2/2(d-1)}.$$

Therefore $\mathbb{P}(B_{r,d}) \leq 2^{-r^2/2(d-1)+\eta r^2}$ as claimed. □

**Theorem 2.8** *For fixed $d \geq 3$, the probability that, in a random subset $S$ of $\mathbf{2}^n$, every point of $S$ lies in a $d$-subcube in $S$ is*

$$\exp\left(-2^{\frac{d-1}{2}\lg^2 n + o(\lg^2 n)}\right).$$

*Therefore*

$$\mathrm{CU}(d; n) = 2^{2^n} \exp\left(-2^{\frac{d-1}{2}\lg^2 n + o(\lg^2 n)}\right).$$

**Proof.** Recall that

$$\mathbb{P}(A_d(x)) = 2^{-(n+1)} \sum_{r=0}^{n} \binom{n}{r} \mathbb{P}(B_{r,d}).$$

For $\eta > 0$, take any $n > 2^{r_0}$, where $r_0 = r_0(\eta)$ is as in the previous lemma. Then we have

$$2^{n+1}\mathbb{P}(A_d(x)) = \sum_{r=0}^{n} \binom{n}{r} \mathbb{P}(B_{r,d} \leq \sum_{r=0}^{r_0} n^r + \sum_{r=r_0}^{n} n^r 2^{-r^2/2(d-1)+\eta r^2}.$$

Each term in the first sum is at most $n^{r_0} \leq 2^{\lg^2 n}$, while the $r$-term $Y_r$ in the second sum is equal to $2^{-\beta r^2 + r \lg n}$, where $\beta = \frac{1}{2(d-1)} - \eta$. Now we have $Y_r \leq 2^{\lg^2 n/4\beta}$, and $1/4\beta < \frac{d-1}{2} + d^2\eta$ for suitably small $\eta$. Therefore

$$\mathbb{P}(A_d(x)) \leq 2^{-(n+1)} n 2^{\frac{d-1}{2}\lg^2 n + d^2\eta \lg^2 n},$$

for any $n > 2^{r_0(\eta)}$.

We also have a lower bound

$$\mathbb{P}(A_d(x)) \geq 2^{-(n+1)} \binom{n}{r_1} 2^{-r_1^2/2(d-1)},$$

where $r_1 = \lceil (d-1)\lg n \rceil$. This gives

$$\mathbb{P}(A_d(x)) \geq 2^{-n+(d-1)\lg^2 n - \frac{(d-1)}{2}\lg^2 n + o(\lg^2 n)}.$$

From these bounds we conclude that

$$\mathbb{P}(A_d(x)) = 2^{-n+\frac{d-1}{2}\lg^2 n + o(\lg^2 n)},$$

as $n \to \infty$.

We also see that, for $n$ sufficiently large, $\mathbb{P}(B_{\lceil n^{3/4}\rceil, d}) \le 2^{-n^{3/2}(\frac{1}{2(d-1)}+\eta)} \le 2^{-2n}$. Having verified this condition, the result now follows from Lemma 2.3. □

For $d$ growing, even very slowly, with $n$, we cannot use Szemerédi's Uniformity Lemma. We suspect that Theorem 2.8 remains valid for $d$ growing more slowly than about $\lg\lg\lg n$ but, as we shall now show, the behaviour is quite different just above that.

For the next result we need a version of the Janson inequalities (see Janson, Luczak and Ruciński [5]). These results are similar to Suen's Inequality, but apply in a more specialised framework. For more details and proofs, see also Alon and Spencer [1].

**Theorem 2.9** *Let $L$ be a set, and let $H$ be a random subset of $L$ defined by taking each element $\ell$ of $L$ with probability $p_\ell$. Let $(L_i) = \{L_i : i \in I\}$ be a family of subsets of $L$, and, for $i \in I$, let $C_i$ be the event that $L_i \subseteq H$. Define a relation $\sim$ on the index set $I$ by $i \sim j$ if $|L_i \cap L_j| \ge 1$ and $i \ne j$. Set*

$$\mu = \sum_{i\in I} \mathbb{P}(C_i); \quad \Delta = \sum_{i\in I}\sum_{j\sim i} \mathbb{P}(C_i \text{ and } C_j).$$

*Then*

$$\prod_{i\in I} \mathbb{P}(\overline{C}_i) \le \mathbb{P}(\text{no } C_i \text{ occurs}) \le e^{-\mu+\Delta/2}.$$

*If also $\Delta \ge \mu$, then*

$$\mathbb{P}(\text{no } C_i \text{ occurs}) \le e^{-\mu^2/2\Delta}.$$

The lower bound in the first inequality above is basically the statement that the events $C_i$ are non-negatively correlated: it is the upper bounds that constitute the Janson Inequalities.

Our next result is part (3) of Theorem 1.2.

**Theorem 2.10** *If $n$ is sufficiently large, and*

$$\lg\lg\lg n + \lg\lg\lg\lg n + 1 \le d \le \lg\lg n + \lg\lg\lg n - 2,$$

*then*

$$2^{2^n} \exp\left(-f(n;d)^2\right) \le \mathrm{CU}(d;n) \le 2^{2^n} \exp\left(-f(n;d)^{1/20}\right),$$

*where*

$$f(n;d) = \exp\left(2^{2^d/(d-1)} (\log n)^{d/(d-1)} d\right).$$

This result shows fairly precisely how $\mathrm{CU}(d;n)$ grows from about $2^{2^n} \exp\left(-2^{\lg^3 n}\right)$ to about $2^{2^n} \exp\left(-2^{n^{1/4}}\right)$. The probability that a random subset $S$ of $\mathbf{2}^n$ is a union of $d$-subcubes falls off very roughly as $2^{-2^{2^{2^d}}}$. (This is comfortably fast enough that the ranges stated above for $\mathrm{CU}(d;n)$, for different $d$, do not intersect.) Given that we know this probability is already as small as $2^{-2^{O(\lg^2 n)}}$ for constant $d$, and that the probability is certainly not smaller than $2^{-2^n}$, the range of $d$ stated above is almost as large as it could be: the inequalities do not hold if the bounds are extended by 2 in either direction.

**Proof.** For convenience, let us note that our assumptions on $d$ imply that:

$$\frac{3}{2}\lg\lg n \le \frac{2^d}{d-1} \le \frac{1}{4}\lg n,$$

so also $\lg^{3/2} n \le 2^{2^d/(d-1)} \le n^{1/4}$, and $\exp(\lg^{5/2} n) \le f(n;d) \le \exp(n^{1/3})$.

Our basic strategy is still to use Lemma 2.3. Accordingly, our main aim is to prove that

$$\frac{1}{2} f(n;d)^{1/18} \le 2^n \mathbb{P}(A_d(x)) \le \frac{n}{2} f(n;d), \tag{1}$$

which will certainly verify the condition $\mathbb{P}(A_d(x)) = 2^{-n+o(n)}$, required for an application of Lemma 2.3. A minor variation on our proof will show that the condition

$$\mathbb{P}(B_{\lceil n^{3/4}\rceil,d}) \le 2^{-2n}$$

is also met for sufficiently large $n$, and so (as we certainly have the final condition $d = o(n)$) the result will follow.

Recall that $2^n \mathbb{P}(A_d(x)) = \frac{1}{2}\sum_{r=0}^{n} \binom{n}{r} \mathbb{P}(B_{r,d})$. To prove 1, we will show that, for some $r_0$,

$$\binom{n}{r_0} \mathbb{P}(B_{r_0,d}) \ge f(n;d)^{1/18},$$

while, for all $r$,

$$\binom{n}{r} \mathbb{P}(B_{r,d}) \le f(n;d).$$

Let $R$ be a fixed $r$-set, let $L$ be the family of subsets of $R$ of size at least 2, and as before let $H$ be a random hypergraph defined by taking each element of $L$ independently with probability $1/2$. For a $d$-set $D$, let $C_D$ be the event that all subsets of $D$ of size at least 2 are in $H$. Thus $\mathbb{P}(C_D) = 2^{-2^d+d+1}$, and $B_{r,d}$ is the event that no $C_D$ occurs. The Janson Inequalities apply to the family $(C_D)$ of events, with $D \sim E$ if $|D \cap E| \ge 2$.

For the lower bound on some $\mathbb{P}(B_{r_0,d})$, we need only that the events $C_D$ are mutually non-negatively correlated (see Theorem 2.9), so

$$\mathbb{P}(B_{r,d}) \ge \left(1 - 2^{-2^d+d+1}\right)^{\binom{r}{d}}$$

and

$$\log \mathbb{P}(B_{r,d}) \geq -\left(\frac{er}{d}\right)^d 2^{-2^d} 3^d.$$

Let $r_0 = \left\lfloor \frac{d}{3e}\left(2^{2^d}\log n\right)^{1/(d-1)} \right\rfloor$. Note that $r_0 \leq 2^{2^d/(d-1)} \log n \leq n^{1/4}\log n$ and so $\log\binom{n}{r_0} \geq r_0 \log(n/r_0) \geq \frac{2}{3} r_0 \log n$. Therefore

$$\log\left(\binom{n}{r_0}\mathbb{P}(B_{r_0,d})\right) \geq r_0\left(\frac{2}{3}\log n - r_0^{d-1}\left(\frac{3e}{d}\right)^d 2^{-2^d}\right) \geq r_0 \log n\left(\frac{2}{3} - \frac{3e}{d}\right) \geq \frac{1}{2} r_0 \log n,$$

since $r_0^{d-1} \leq 2^{2^d}\log n (d/3e)^{d-1}$. This gives that

$$\binom{n}{r_0}\mathbb{P}(B_{r_0,d}) \geq \exp\left(2^{2^d/(d-1)}(\log n)^{d/(d-1)}\frac{d}{18}\right) = f(n;d)^{1/18},$$

as claimed.

It remains to prove upper bounds on $\mathbb{P}(B_{r,d})$, namely that $\binom{n}{r}\mathbb{P}(B_{r,d}) \leq f(n;d)$ for all $r$, and specifically $\mathbb{P}(B_{\lceil n^{3/4}\rceil,d}) \leq 2^{-2n}$, for sufficiently large $n$. Let $r_1 = 2^{2^d/(d-1)}(\log n)^{1/(d-1)}d$, and note that $r_1 \geq \log^{3/2} n$. For $r \leq r_1$, we have $\binom{n}{r} \leq n^{r_1} \leq \exp\left\{2^{2^d/(d-1)}(\log n)^{d/(d-1)}d\right\} = f(n;d)$, which is as required. For $r > r_1$, we claim that $n^r\mathbb{P}(B_{r,d}) \leq 1$, or equivalently $-\log \mathbb{P}(B_{r,d}) \geq r \log n$, which will imply that $\binom{n}{r}\mathbb{P}(B_{r,d}) \leq 1 \leq f(n;d)$.

As mentioned earlier, we can apply the Janson Inequalities, Theorem 2.9, to the events $C_D$. The quantities appearing in the Inequalities are

$$\mu = \sum_D \mathbb{P}(C_D) = \binom{r}{d} 2^{-2^d+d+1}$$

and

$$\Delta = \sum_D \sum_{E \sim D} \mathbb{P}(C_D \text{ and } C_E) = \binom{r}{d}\sum_{\ell=2}^{d-1}\binom{d}{\ell}\binom{r-d}{d-\ell}2^{-2\cdot 2^d + 2^\ell + 2d - \ell + 1} = \binom{r}{d}\sum_{\ell=2}^{d-1} U_\ell,$$

where the $\ell$-term $U_\ell$ in the sum accounts for the $d$-sets $E$ intersecting $D$ in exactly $\ell$ elements. A short calculation shows that the sequence $(U_\ell)$ of terms is log-convex, so the largest term is either the first or the last. Hence

$$\Delta \leq \binom{r}{d}(d-2)\max\left\{\binom{d}{2}\binom{r-d}{d-2}2^{-2\cdot 2^d + 2d + 3}, d(r-d)2^{-\frac{3}{2}2^d + d + 2}\right\}.$$

Since $\binom{r-d}{d-2}/\binom{r}{d} \leq d^2/r^2$, we can write this as

$$\Delta \leq \max\{\frac{\mu^2 d^5}{r^2}, 2\mu d^2 r 2^{-\frac{1}{2}2^d}\}.$$

As it happens, for $r > r_1$ the first term is always larger; to see this, recall that $\mu = \binom{r}{d}2^{-2^d+d+1} \ge \left(\frac{r}{d}\right)^d 2^{-2^d}$, and observe that

$$\frac{\mu d^3 2^{\frac{1}{2}2^d}}{r^3} \ge \left(\frac{r}{d}\right)^{d-3} 2^{-\frac{1}{2}2^d} \ge \left(\frac{r}{d}\right)^{d-3} r^{-(d-1)/2} \ge 1.$$

Hence $\Delta \le \mu^2 d^5/r^2$. Depending on the value of $r$, it may or may not be the case that $\Delta \le \mu$, so we shall need both Janson Inequalities. Together, they tell us that

$$\mathbb{P}(B_{r,d}) = \mathbb{P}(\text{no } C_D \text{ occurs}) \le \max\{e^{-\mu/2}, e^{-\mu^2/2\Delta}\}.$$

Combining this with our bound $\Delta \le \mu^2 d^5/r^2$ gives

$$-\log \mathbb{P}(B_{r,d}) \ge \min\{\mu/2, r^2/2d^5\}.$$

Now

$$\frac{\mu}{2} \ge \left(\frac{2r}{d}\right)^d 2^{-2^d} \ge r \log n \frac{2^d}{d} \ge r \log n,$$

since $r^{d-1} \ge (r_1+1)^{d-1} \ge 2^{2^d} \log n d^{d-1}$. Also $r/2d^5 \ge \log^{3/2} n/2d^5 \ge \log n$, so we have $-\log \mathbb{P}(B_{r,d}) \ge r \log n$ for all $r > r_1$, as required. This establishes equation 1.

To complete the proof, we need to check that $-\log \mathbb{P}(B_{\lceil n^{3/4}\rceil,d}) \ge 2n$, for $n$ sufficiently large. This follows from the above calculations since, for $r = \lceil n^{3/4}\rceil$ and $n$ sufficiently large,

$$\frac{\mu}{2} \ge \left(\frac{2\lceil n^{3/4}\rceil}{d}\right)^d 2^{-2^d} \ge \left(\frac{2n^{3/4}}{d}\right)^d n^{-\frac{1}{4}(d-1)} \ge n^{d/4}$$

and $r^2/2d^5 \ge n^{3/2}/\log n$. □

Consider the quantity $X(d;n) = 2^n - \lg \mathrm{CU}(d;n)$ (this is the logarithm of the reciprocal of the probability that a random set is a union of $d$-subcubes). For constant values of $d$, $X(d;n)$ is around $2^{\frac{d-1}{2}\lg^2 n}$, while for values of $d$ above about $\sqrt{\lg n}$, $X(d;n)$ is already much larger, namely $2^n$ divided by a term polynomial in $n/d$. Theorem 2.10 shows how $X(d;n)$ grows very rapidly from about $2^{\lg^{5/2} n}$ to about $2^{n^{1/4}}$. We leave open the question of what happens just either side of the range covered in Theorem 2.10: it seems reasonable to expect that $X(d;n) \sim 2^{\lg^2 n}$ for $d \le \lg\lg\lg n$, and that $X(d;n) = 2^{n^{1-o(1)}}$ for $d \ge (1+\varepsilon)\lg\lg n$.

# 3 Lower Bounds

We now move on to higher values of $d$. Here to get a lower bound we have to construct many unions of $d$-subcubes, rather than hope to find one by chance.

We first state our result in a general form, and then specialise to various ranges of $d$. Not surprisingly, this result is not close to the truth for $d \le \lg\lg n$.

**Theorem 3.1** *For any function $d = d(n)$ taking values in the positive integers and satisfying $(n-2d)/\sqrt{n} \to \infty$, we have*

$$\lg \mathrm{CU}(d;n) \geq 2^n \frac{n-2d}{n-d} \left(\frac{d}{n-d}\right)^{d/(n-2d)(1+o(1))},$$

*as $n \to \infty$.*

**Proof.** Take any set $T \subseteq \mathbf{2}^n$, and let $U = U(T;d)$ be the set of points at distance exactly $d$ from $T$. Now, for each point $u \in U$, take a $d$-subcube $C_u$ including both $u$ and a point $t_u$ of $T$; necessarily $C_u \cap U = \{u\}$, since all other points of $C_u$ are at distance less than $d$ from the point $t_u$ of $T$. Now, for each subset $V$ of $U$, consider the set $R_V = \bigcup_{u \in V} C_u$. All the sets $R_V$ are distinct, since $R_V \cap U = V$ for each $V$. Therefore $\mathrm{CU}(d;n) \geq 2^{|U|}$ for any set $U = U(T;d)$. To complete the proof, we will show that there is some set $T$ such that

$$|U(T;d)| \geq 2^n \frac{n-2d}{n-d} \left(\frac{d}{n-d}\right)^{\frac{d}{n-2d}(1+o(1))}.$$

Let $T$ be a set chosen at random, with $\mathbb{P}(y \in T) = p$ for all $y$, all choices made independently; $p$ will be specified shortly. The probability that a fixed $x \in \mathbf{2}^n$ is at distance exactly $d$ from such a random set $T$ is then

$$(1-p)^{1+n+\cdots+\binom{n}{d-2}+\binom{n}{d-1}} \left(1-(1-p)^{\binom{n}{d}}\right).$$

Now, provided $(n-2d)/\sqrt{n} \to \infty$, we have

$$1+n+\cdots+\binom{n}{d-2}+\binom{n}{d-1} = \frac{d}{n-2d}\binom{n}{d}(1+o(1)),$$

and so

$$\mathbb{P}(d(x,T)=d) = (1-p)^{\frac{d}{n-2d}\binom{n}{d}(1+o(1))} \left(1-(1-p)^{\binom{n}{d}}\right).$$

To maximise the above probability, we choose $p$ so that $(1-p)^{\binom{n}{d}} = \frac{d}{n-d}$, and so obtain

$$\mathbb{P}(d(x,T)=d) = \frac{n-2d}{n-d} \left(\frac{d}{n-d}\right)^{\frac{d}{n-2d}(1+o(1))}.$$

This gives the desired result, since $\mathbb{E}|U(T;d)| \geq 2^n \mathbb{P}(d(x,T)=d)$. □

The bound in Theorem 3.1 can be made more explicit in various ranges covered by the theorem. We obtain the following:

$$\lg \mathrm{CU}(d;n) \geq 2^n \left(1 - \frac{d \log(n/d)}{n}(1+o(1))\right) \qquad \text{for } d = o(n);$$

$$\lg \mathrm{CU}(d;n) \geq \beta 2^n (1+o(1)) \qquad \text{for } d = \alpha n \text{ with } 0 < \alpha < 1/2,$$

where

$$\beta = \frac{1-2\alpha}{1-\alpha}\left(\frac{\alpha}{1-\alpha}\right)^{\alpha/(1-2\alpha)};$$

$$\lg \mathrm{CU}(d;n) \geq 2^n \binom{4m}{en}(1+o(1)) \qquad \text{for } d = n/2 - m, \text{ with } \sqrt{n} \ll m \ll n.$$

For $d = (\frac{1}{2} - \varepsilon)n$, we have that $\lg \mathrm{CU}(d;n) \geq \delta 2^n$, for some $\delta = \delta(\varepsilon) > 0$. On the other hand, part (6) of Theorem 1.2, from [3], implies that, for $d = (\frac{1}{2} + \varepsilon)n$, $\lg \mathrm{CU}(d;n) \leq 2^{(1-\delta)n}$ for some $\delta = \delta(\varepsilon) > 0$. Thus there is a definite "phase transition" around $d = n/2$, and the proof of Theorem 3.1 perhaps gives some insight into why this is the case. Indeed, if we try to use the same construction as in that proof when $d \geq n/2$, the best we can do is to take $|T| = 1$, which gives only $\mathrm{CU}(d;n) \geq 2^{\binom{n}{d}}$. Of course, this is the same bound as we obtain from the family of monotone formulae, because the two constructions amount to the same thing. As mentioned earlier, we think that this simple lower bound may be roughly correct, at least for $d \geq (\frac{1}{2} + \varepsilon)n$.

# 4 The Upper Bound

Our purpose in this section is to establish the upper bound stated in parts (4) and (5) of Theorem 1.2.

Our result will be proved using techniques of entropy, as presented by Kahn [6, 7]. Our method will follow that used in [6], in particular in the proof of Theorem 1.9, pp.226-7. Our description of the method will be somewhat brief; we encourage the interested reader to consult Kahn's papers.

Let $\mathbf{X}$ be a discrete random variable, taking values in a finite set $J$, with $\mathbb{P}(\mathbf{X} = j) = p_j$ for $j \in J$. The *entropy* of $\mathbf{X}$ is

$$H(\mathbf{X}) = \sum_{j \in J} p_j \log(1/p_j).$$

We have $H(\mathbf{X}) \leq \lg |J|$, with equality if and only if $\mathbf{X}$ is uniform on $J$.

If $\mathbf{Y}$ is another discrete random variable, then the *conditional entropy* of $\mathbf{X}$ given $\mathbf{Y}$ is

$$H(\mathbf{X} \mid \mathbf{Y}) = \mathbb{E}H(\mathbf{X} \mid \mathbf{Y} = y).$$

If $\mathbf{Z}$ is a function of $\mathbf{Y}$, then $H(\mathbf{X} \mid \mathbf{Y}) \leq H(\mathbf{X} \mid \mathbf{Z})$.

If $\mathbf{X}_1, \ldots, \mathbf{X}_m$ are discrete random variables on the same probability space, then we can regard the random vector $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_m)$ as a single discrete random variable, and we have

$$H(\mathbf{X}) = H(\mathbf{X}_1) + H(\mathbf{X}_2 \mid \mathbf{X}_1) + \cdots + H(\mathbf{X}_m \mid \mathbf{X}_1, \ldots, \mathbf{X}_{m-1}),$$

which implies that

$$H(\mathbf{X}) \leq \sum_{i=1}^{m} H(\mathbf{X}_i).$$

A powerful extension of this last inequality is the following result, proved by Shearer (see [4]), and stated in this form by Kahn [6, 7]. If $\mathbf{X}_1, \ldots, \mathbf{X}_m$ are random variables on the same probability space, and $A = \{a_1, \ldots, a_s\} \subseteq [m]$, let $\mathbf{X}_A$ be the random vector $(\mathbf{X}_{a_1}, \ldots, \mathbf{X}_{a_s})$. Shearer's Lemma is as follows.

**Lemma 4.1** *Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_m)$ be a random vector, and let $\mathcal{A} = (A_1, \ldots, A_t)$ be a family of subsets of $[m]$ (possibly with repetitions), with each element of $[m]$ covered at least $k$ times by the $A_\ell$. Then*

$$H(\mathbf{X}) \leq \frac{1}{k} \sum_{\ell=1}^{t} H(\mathbf{X}_{A_\ell}).$$

We are now ready for our main upper bound on $\mathrm{CU}(d; n)$.

**Theorem 4.2** *For any $d \geq 18$ and $n$ with $6\sqrt{\lg n} \leq d \leq n/2$,*

$$\lg \mathrm{CU}(d; n) \leq 2^n \left(1 - \frac{d^3}{3n^3}\right).$$

**Proof.** Let $\mathcal{E}$ and $\mathcal{O}$ be the two classes of the bipartition of the cube $\mathbf{2}^n$.

The general principle underlying the proof is that, if some point $x$ is in a union $S$ of $d$-subcubes, then this severely restricts the choice of what can happen to the set of points at distance 2 from $x$; in particular there are at most $2^{\binom{n}{2}} \binom{n}{d} 2^{-\binom{d}{2}}$ choices for how $S$ can intersect this set. We are not able to use this idea as it stands; however if we move to the set of points at distance 3 from $x$, then we can see explicitly how choices of $S \cap \mathcal{E}$ severely constrain how we can choose $S \cap \mathcal{O}$. The entropy framework, in particular Shearer's Lemma, allows us to bound how much "information" the random set $S$ encodes.

We continue to follow Kahn [6, 7], in particular pp.226-7 of [6]. Let $\mathbf{S}$ be a set chosen uniformly at random from those subsets of $\mathbf{2}^n$ that are unions of $d$-subcubes. We think of $\mathbf{S}$ as a discrete random variable, so that $H(\mathbf{S}) = \lg \mathrm{CU}(d; n)$.

For $x \in \mathbf{2}^n$, let $\mathbf{1}_x$ denote the indicator random variable of the event that $x$ is in $\mathbf{S}$. Let $\mathbf{X}_x = (\mathbf{1}_y : y \in N^3(x))$, where $N^3(x)$ is the set of points at distance 3 from $x$ in $\mathbf{2}^n$. Thus $\mathbf{X}_x$ encodes the intersection of $\mathbf{S}$ with $N^3(x)$.

Let $Q_x$ be the event that, for some $d$-subcube $C$ containing $x$, $C \cap N^3(x) \subseteq \mathbf{S}$. By symmetry, the probability of $Q_x$ in this uniform measure is independent of $x$ – we denote the probability by $q$.

As in the papers of Kahn, we have that

$$\lg \mathrm{CU}(d; n) = H(\mathbf{S}) = H(\mathbf{S} \cap \mathcal{O}) + H(\mathbf{S} \cap \mathcal{E} \mid \mathbf{S} \cap \mathcal{O}).$$

Now

$$H(\mathbf{S}\cap\mathcal{E}\mid\mathbf{S}\cap\mathcal{O})\le\sum_{x\in\mathcal{E}}H(\mathbf{1}_x\mid\mathbf{S}\cap\mathcal{O})\le\sum_{x\in\mathcal{E}}H(\mathbf{1}_x\mid\mathbf{1}_{Q_x})=2^{n-1}H(\mathbf{1}_v\mid\mathbf{1}_{Q_v}),$$

for any $v\in\mathcal{E}$, where the equality holds since all the terms are equal. Furthermore,

$$H(\mathbf{1}_v|\mathbf{1}_{Q_v})=qH(\mathbf{1}_v|Q_v)+(1-q)H(\mathbf{1}_v|\overline{Q_v})\le q\times 1+(1-q)\times 0=q,$$

since the event $Q_x$ is necessary for $x$ to be in $\mathbf{S}$. Therefore we have

$$H(\mathbf{S}\cap\mathcal{E}\mid\mathbf{S}\cap\mathcal{O})\le 2^{n-1}q.$$

To bound $H(\mathbf{S}\cap\mathcal{O})$, we apply Shearer's Lemma with the family $\{N^3(x):x\in\mathcal{E}\}$, noting that each element $y\in\mathcal{O}$ is covered exactly $\binom{n}{3}$ times by the members of this family. We obtain

$$\begin{aligned}H(\mathbf{S}\cap\mathcal{O})&\le\tfrac{1}{\binom{n}{3}}\textstyle\sum_{x\in\mathcal{E}}H(\mathbf{X}_x)\\&=\tfrac{2^{n-1}}{\binom{n}{3}}\left(H(\mathbf{1}_{Q_v})+H(\mathbf{X}_v\mid\mathbf{1}_{Q_v})\right),\end{aligned}$$

for any $v\in\mathcal{E}$. Certainly $H(\mathbf{1}_{Q_v})\le 1$, while

$$H(\mathbf{X}_v|\mathbf{1}_{Q_v})=qH(\mathbf{X}_v|Q_v)+(1-q)H(\mathbf{X}_v|\overline{Q_v}).$$

We see that $H(\mathbf{X}_v|\overline{Q_v})\le|N^3(v)|=\binom{n}{3}$. Under $Q_v$, $\mathbf{X}_v$ can take at most $2^{\binom{n}{3}}\binom{n}{d}2^{-\binom{d}{3}}$ values, so

$$H(\mathbf{X}_v|Q_v)\le\lg\left(2^{\binom{n}{3}}\binom{n}{d}2^{-\binom{d}{3}}\right)\le\binom{n}{3}-\binom{d}{3}+d\lg n.$$

Now we have

$$H(\mathbf{S}\cap\mathcal{O})\le\frac{2^{n-1}}{\binom{n}{3}}\left(1+q\left\{\binom{n}{3}-\binom{d}{3}+d\lg n\right\}+(1-q)\binom{n}{3}\right).$$

Combining the bounds on $H(\mathbf{S}\cap\mathcal{E}\mid\mathbf{S}\cap\mathcal{O})$ and on $H(\mathbf{S}\cap\mathcal{O})$ gives:

$$\begin{aligned}\lg\mathrm{CU}(d;n)&\le\tfrac{2^{n-1}}{\binom{n}{3}}\left[1+q\left\{\binom{n}{3}-\binom{d}{3}+d\lg n\right\}+(1-q)\binom{n}{3}\right]+2^{n-1}q\\&=2^{n-1}\left[1+\tfrac{1}{\binom{n}{3}}+q\left(1-\tfrac{\binom{d}{3}-d\lg n}{\binom{n}{3}}\right)\right].\end{aligned}$$

Certainly $q\le 1$, so

$$\lg\mathrm{CU}(d;n)\le 2^{n-1}\left[2-\frac{d(d-1)(d-2)-1-6d\lg n}{n^3}\right]\le 2^n\left[1-\frac{5d^3/6-d^3/6}{2n^3}\right],$$

which is the result claimed. □

While the bounds from the last two sections – see parts (4) and (5) of Theorem 1.2 – reveal the basic behaviour of $\mathrm{CU}(d;n)$, they are still quite far apart. For instance, if $d=n/4$, our results are that

$$\frac{2}{3\sqrt{3}}(1+o(1))\le 2^{-n}\lg\mathrm{CU}(d;n)\le 1-\frac{1}{192}.$$

To be specific, one topic that seems worthy of study is to determine how fast $2^{-n}\lg \mathrm{CU}(d;n)$ approaches 1 as $d/n$ tends to 0 slowly. Our upper bound of $1-O((d/n)^3)$ on this function seems unlikely to be correct, but we think our lower bound of $1-O((d/n)\log(n/d))$ might not be far from the truth.

## 5 Convergence

In this section, we change tack somewhat. We concentrate on a particular range of $d$, namely $d=\alpha n$, for $\alpha \leq 1/2$ constant. In this range, we have shown that

$$2^{\beta'(\alpha)2^n(1+o(1))} \leq \mathrm{CU}(\alpha n;n) \leq 2^{\beta''(\alpha)2^n(1+o(1))},$$

where $0<\beta'(\alpha)$ and $\beta''(\alpha)<1$ for $\alpha \in (0,1/2]$. It seems very likely that there is a single function $\beta(\alpha)$ such that $\mathrm{CU}(\alpha n;n)=2^{\beta(\alpha)2^n(1+o(1))}$ as $n\to\infty$. We shall not quite be able to prove this, but we shall get very close.

To state our result precisely, we define

$$\beta_n(\alpha)=\frac{\lg \mathrm{CU}(\alpha n;n)}{2^n},$$

for every $n$ and every $\alpha \in (0,\frac{1}{2}]$.

One natural approach to proving that $\beta_n(\alpha)$ converges for each fixed $\alpha$ would be to show that the function is monotonic in $n$. We do not know whether this is true, but the next result is of a similar nature.

**Lemma 5.1** *There exists an absolute constant $n_0$ such that, for any $\alpha \in (0,\frac{1}{2}]$, and any integers $n$, $m$ with $n_0 \leq n \leq m \leq 2n$,*

$$\beta_m(\alpha) \leq \beta_n(\alpha - \log n/\sqrt{n}) + e^{-2\log^2 n}.$$

**Proof.** Let $S$ be any subset of $\mathbf{2}^m$ that is a union of cubes of dimension at least $\alpha m$. Such a set can be written as a union of at most $2^m$ such subcubes; say $S=C_1\cup\cdots\cup C_t$. Suppose that $C_j$ is defined by fixing the co-ordinates outside the set $V_j\subset[m]$, so $|V_j|\geq \alpha m$ for $j=1,\dots,t$.

Now choose a random set $A\subset[m]$ with $|A|=n$. For each $j$, the probability that $|A\cap V_j|$ is less than $\alpha n-\sqrt{n}\log n$ is at most $\sum_{k\geq\sqrt{n}\log n}\left(1+\frac{2k}{n}\right)^{-2k}$, for any $\alpha\in(0,\frac{1}{2})$ and any $m\in[n,2n]$. This probability is at most $e^{-3\log^2 n}$, for sufficiently large $n$. Therefore there is a set $A\subset[m]$ of size $n$ such that $|A\cap V_j|\geq \alpha n-\sqrt{n}\log n$ for all but at most $2^m e^{-3\log^2 n}$ of the subcubes $C_j$.

For such an $A$, we decompose $\mathbf{2}^m$ as the union of $s=2^{m-n}$ cubes $D_1,\dots,D_s$ of dimension $n$, each defined by a choice of the co-ordinates outside $A$. If $C_j$ intersects $D_i$, then their intersection is a subcube of dimension $|V_j\cap A|$.

Hence, for some $A$, $S$ can be written as a union $S_0 \cup S_1 \cup \cdots \cup S_s$, where $S_0$ is a union of at most $2^m e^{-3\log^2 n}$ subcubes, and $S_i \subseteq D_i$ is a union of cubes each of dimension at least $\alpha n - \sqrt{n}\log n$.

The number of choices for $A$ is at most $2^m$. The number of choices for $S_0$ is at most $(3^m)^{2^m e^{-3\log^2 n}}$, since the total number of subcubes of $\mathbf{2}^m$ is $3^m$. The number of choices for each $S_i$ is $\mathrm{CU}(\alpha n - \sqrt{n}\log n; n) = 2^{2^n \beta_n(\alpha - \log n/\sqrt{n})}$. Therefore the binary logarithm of the number $\mathrm{CU}(\alpha m; m)$ of possible sets $S$ is at most

$$m + \lg 3m2^m e^{-3\log^2 n} + s2^n \beta_n(\alpha - \log n/\sqrt{n}) \leq 2^m \left(\beta_n(\alpha - \log n/\sqrt{n}) + e^{-2\log^2 n}\right).$$

By definition of $\beta_m(\alpha)$, this is the claimed result. □

We now prove our main result of this section.

**Theorem 5.2** *Suppose* $0 \leq \alpha_1 < \alpha_2 \leq \frac{1}{2}$. *Then*

$$\liminf_n \beta_n(\alpha_1) \geq \limsup_n \beta_n(\alpha_2).$$

**Proof.** Suppose the statement is false, and take any $\alpha_1 < \alpha_2$, and $\varepsilon > 0$, such that $\liminf \beta_n(\alpha_1) = \beta_0$ while $\limsup \beta_n(\alpha_2) \geq \beta_0 + \varepsilon$.

Choose $n_1$ larger than the constant $n_0$ of the previous lemma, such that in addition: $\beta_{n_1}(\alpha_1) \leq \beta_0 + \varepsilon/3$, $\sum_{k=0}^{\infty} e^{-2\log^2(n_1 2^k)} \leq \varepsilon/3$, and $\sum_{k=0}^{\infty} \frac{\log(n_1 2^k)}{\sqrt{n_1 2^k}} \leq \alpha_2 - \alpha_1$.

Now take any $n' \geq n_1$, and let $t$ be the integer such that $2^t n_1 \leq n' < 2^{t+1} n_1$. We claim that, for $\ell = 0, \ldots, t$,

$$\beta_{n_1 2^\ell}\left(\alpha_1 + \sum_{k=0}^{\ell-1} \frac{\log(n_1 2^k)}{\sqrt{n_1 2^k}}\right) \leq \beta_{n_1}(\alpha_1) + \sum_{k=0}^{\ell-1} e^{-2\log^2(n_1 2^k)}.$$

Indeed, the statement for $\ell = 0$ is trivial, and the statement for $\ell > 0$ follows by induction on using the previous lemma, setting $n = n_1 2^{\ell-1}$ and $m = n_1 2^\ell$. Applying the previous lemma again then shows that

$$\beta_{n'}(\alpha_2) \leq \beta_{n'}\left(\alpha_1 + \sum_{k=0}^{\ell} \frac{\log(n_1 2^k)}{\sqrt{n_1 2^k}}\right) \leq \beta_{n_1}(\alpha_1) + \sum_{k=0}^{\ell} e^{-2\log^2(n_1 2^k)} \leq \beta_0 + 2\varepsilon/3.$$

This holds for all $n' \geq n_1$, contradicting the assumption that $\limsup \beta_n(\alpha_2) \geq \beta_0 + \varepsilon$. □

What this tells us is that, for any $\beta$, there is at most one $\alpha$ such that $\liminf \beta_n(\alpha) < \beta < \limsup \beta_n(\alpha)$. In particular this implies that there are at most countably many $\alpha$ such that $\lim_{n\to\infty} \beta_n(\alpha)$ does not exist. The following obvious conjecture is surely true.

**Conjecture 5.3** *For every* $\alpha \in (0, \frac{1}{2}]$, *there is a real* $\beta = \beta(\alpha)$ *such that* $\beta_n(\alpha) \to \beta$ *as* $n \to \infty$. *Equivalently,* $\mathrm{CU}(\alpha n; n) = 2^{\beta 2^n(1+o(1))}$ *as* $n \to \infty$.

We know that, for $0 < \alpha < 1/2$, the value $\beta(\alpha)$ would have to lie strictly between 0 and 1.

# 6 Open Problems

While it would be nice to settle Conjecture 5.3, we feel that there are more important issues left to resolve, and we conclude by reminding the reader of some of the main outstanding problems.

As we have already mentioned, there are some ranges of $d = d(n)$ not covered by Theorem 1.2, and there are significant gaps between the lower and upper bounds in several parts of that result, notably (5), dealing with the range $d = \alpha n$, with $\alpha < 1/2$.

We also wish to stress that the picture for $d = n - k$, $k$ constant, is far from complete. In particular, Bollobás, Brightwell and Leader [3] showed that there are constants $c_k$ with $\lg \mathrm{SAT}(k; n) = c_k \binom{n}{k}(1 + o(1))$ for each fixed $k \geq 2$, but were only able to determine $c_k$ for $k = 2$. Our Conjecture 1.1 would imply that $c_k = 1$ for all $k$.

# References

[1] Noga Alon and Joel H. Spencer, *The Probabilistic Method*, 2nd Edition, Wiley Interscience (2000), xvi+301pp.

[2] Béla Bollobás, *Modern Graph Theory*, Springer-Verlag 1998, xiii+394pp.

[3] Béla Bollobás, Graham Brightwell and Imre Leader, The number of 2-SAT functions, submitted.

[4] F.R.K. Chung, P. Frankl, R. Graham and J.B. Shearer, Some intersection theorems for ordered sets and graphs, *J. Combin. Th. Ser. A* **48** (1986) 23–37.

[5] S. Janson, T. Luczak and A. Ruciński, An exponential bound for the probability of nonexistence of a specified subgraph in a random graph, in *Random Graphs '87*, Proceedings, Poznań 1987, M. Karoński, J. Jaworski and A. Ruciński Eds, John Wiley and Sons, Chichester (1990) 73–87.

[6] Jeff Kahn, An entropy approach to the hard-core model on bipartite graphs, *Combin. Prob. Comput.* **10** (2001) 219–237.

[7] Jeff Kahn, Entropy, independent sets and antichains, manuscript.

[8] W.C. Suen, A correlation inequality and a Poisson limit theorem for nonoverlapping balanced subgraphs of a random graph, *Rand. Struct. Alg.* **1** (1990) 231–242.