Proseminar Theoretische Informatik 19.12.2017

CNF-SAT and other random graphs

Lavinia Kulawik *Wolfgang Mulzer*

1 CNF-SAT

Conjunctive normal form (CNF) formula over the variables $\{x_1, x_2, x_3, x_4\}$:

$$\underbrace{(x_1 \vee \overline{x_2} \vee x_3)}_{\text{clause}} \wedge (x_2 \vee \overline{x_4}) \wedge (\underbrace{x_1}_{\text{literal}} \vee x_4) \wedge (x_3 \vee x_4) \wedge (x_2 \vee \overline{x_3} \vee x_4)$$

k-CNF: each clause has size at most k
CNF-SAT problem: whether a given CNF formula ist satisfiable. It is NP-Complete, any problem in the class NP can be converted into it and there is no efficient algorithm for worst-case instances.

1.1 SAT-solvers in practice

complete solvers

- are guaranteed to find a satisfying assignment whenever one exists.
- DPLL (David-Putnam-Logemann-Loveland) style procedure is an example

incomplete solvers

- may fail to output a solution even when a formula is satisfiable
- typically much faster than complete solvers
- Walksat local-search procedure is an example

1.2 Phase Transitions for CNF-SAT

Generating a random CNF formula f with n variables, m clauses and k literals:

1. Select m clauses independently at random from a set of $\binom{n}{k}2^k$ possible clauses
2. Choose a random set of k distinct variables. Choose to either negate it or not

Adding more clauses (increasing m) preserves unsatisfiability. We conjecture that $r_k n$ is a sharp threshold, r_k is a constant independent of n.

Finding upper bound on r_k

Let $m = cn$. $2^n(1 - \frac{1}{2^k})^{cn}$ is the expected number of satisfying assignments for a formula with cn clauses. If $c = 2^k \ln 2$, $2^n(1 - \frac{1}{2^k})^{n2^k \ln 2} \leq 2^n e^{-n \ln 2} = 2^n 2^{-n} = 1 \quad \Rightarrow r_k \leq 2^k \ln 2$

Finding lower bound on r_k

We focus only on the case $k = 3$. The second moment method cannot be directly applied. With the Smallest Clause Heuristic we can find some lower bound on r_3.
Smallest Clause Heuristic, Unit Clause Heuristic and "pure literal" heuristic set a particular literal w to true, delete all clauses containing w and $\overline{w}$ is deleted from any clause containing $\overline{w}$. If a clause is reduced to length zero, then the algorithm has failed to find a satisfying assignment, but the formula may be satisfiable.

Theorem 1.1. *If the number of clauses in a random 3-CNF formula grows as cn where c is a constant less than $\frac{2}{3}$, then with probability $1-o(1)$, the Shortest Clause Heuristic finds a satisfying assignment.*

Idea of proof 1.1. The arrival rate is the expected number of arrivals into the queue at a given time t. Here it is $\approx \frac{3c}{2}$, which for $c < \frac{2}{3}$ is a constant less than one. This and Lemma 2 and Lemma 4 imply that the probability of two contradictory 1-length clauses, which is a precursor to a 0-length clause, is very small.

Lemma 1.2. *Let T_i be the first time that clause i turns into a 2-clause. T_i is ∞ if clause i gets satisfied before turning into a 2-clause. The T_i are mutually independent over the randomness in constructing the formula and the randomness in SC, and for any t,*

$$Prob(T_i = t) \leq \frac{3}{2(n-2)}.$$

Idea of proof 1.2. Generate the clauses with the method of "deferred decisions" that results in the same distribution as the original. $T_i = t$ if and only if the method of deferred decisions does not put the current literal at steps $1, 2, \ldots, t-1$ into the i^{th} clause, but puts the negation of the literal at step t into it. The probability is precisely

$$\frac{1}{2}(1-\frac{3}{n})(1-\frac{3}{n-1})\ldots(1-\frac{3}{n-t+2})\frac{3}{n-t+1} \leq \frac{3}{2(n-2)}$$

Clearly the T_i are independent since again deferred decisions deal with different clauses independently.

Lemma 1.3. *There exists a constant c_2 such that with probability $1-o(1)$, no clause remains a 2 or 1-clause for more than $c_2 \ln n$ steps. I.e., once a 3-clause becomes a 2-clause, it is either satisfied or reduced to a 0-clause in $O(\ln n)$ steps.*

Idea of proof 1.3. Say that t is a "busy time" if there exists at least one 2-clause or 1-clause at time t, and define a time-window $[r+1, s]$ to be a "busy window" if time r is not busy but then each $t \in [r+1, s]$ is a busy time. We will prove, that for some constant c_2, with probability $1-o(1)$, all busy windows have legth at most $c_2 \ln n$.

$$Prob(\text{a 3-clause turns into a 2-clause during } [r,s]) \leq \frac{3(s-r)}{2(n-2)}.$$

The arrival rate onto the queue of 2 an 1-clauses os a constant strictly less than one. If $s - r \geq c_2 \ln n$ the probability that more than $s-r$ clauses turn into 2-clauses between r and s is at most $1/n^3$. Applying the union bound over all $O(n^2)$ possible choices of r and s, we get that the probability that any clause remains a 2 or 1-clause for more than $c_2 \ln n$ steps is $o(1)$.

2 Nonuniform Models of Random Graphs

Random Graphs with power law degree distribution simulate many real world problems. Number $f(d)$ of vertices of degree d scales as $1/d^\alpha$ for some constant $\alpha > 0$.
Edges are not independent and this makes it harder to analyze, e.g. phase transitions.

2.1 Giant Component in Graphs with Given Degree Distribution

Let λ_i be the fraction of vertices of degree i. There will be a giant component if and only if $\sum_{i=0}^{\infty} i(i-2)\lambda_i > 0$.

3 Growth Models

In Growth Models, vertices and edges are added to the graph over time.
Selecting vertices for attaching a new edge:

- select two vertices uniformly at random from the set of existing vertices.
- select two vertices with probability proportional to their degree (preferential attachment)
- add new vertex at each unit of time and with probability δ add an edge where one end of the edge is the new vertex and the other end is a vertex selected with probability proportional to its degree (power law degree distribution)

3.1 Growth Model Without Preferential Attachment

Multi-graphs possible. Since at time t, there are t vertices and in expectation only $O(\delta t)$ edges where there are t^2 pairs of vertices, it is very unlikely that there will be many multiple edges. Let $d_k(t)$ be the expectation of the number of vertices of degree k at time t. $2\delta\frac{d_0(t)}{t}$ is the expectiation of the number of isolated vertices that are picked to be end points of the new edge. Thus,

$$d_0(t+1) = d_0(t) + 1 - 2\delta\tfrac{d_0(t)}{t},$$
$$d_k(t+1) = d_k(t) + 2\delta\tfrac{d_{k-1}(t)}{t} - 2\delta\tfrac{d_k(t)}{t}$$

There is a solution of the form $d_k(t) = p_k t$, where p_k depends only on k and not on t, provided k is fixed and $t \to \infty$.

$$p_0 = \tfrac{1}{1+2\delta} \text{ and } p_k = \tfrac{2\delta}{1+2\delta}(\tfrac{2\delta}{1+2\delta})^k$$

Thus, the model gives rise to a graph with a degree destribution that falls off exponentially fast with the degree.

The generating function for component size

Let $n_k(t)$ be the expected number of components of size k at time t. We will show that there is a solution for $n_k(t)$ of the form $a_k t$ where a_k is a constant independent of t. The probability that the size of the component containing a rendomly selected vertex in k is proportional to $kn_k(t)$.

$$n_1(t+1) = n_1(t) + 1 - 2\delta\tfrac{n_1(t)}{t}$$

For $k > 1$:

$$n_k(t+1) = n_k(t) + \delta\textstyle\sum_{j=1}^{k-1}\tfrac{jn_j(t)}{t}\tfrac{(k-j)n_{k-j}(t)}{t} - 2\delta\tfrac{kn_k(t)}{t}$$

ka_k is the probability that a random vertex is in a connected component of size k. Solving for a_k yields $a_1 = \frac{1}{1+2\delta}$ and $a_k = \frac{\delta}{1+2k\delta}\sum_{j=1}^{k-1} j(k-j)a_j a_{k-j}$.
Now we use the generating function $g(x)$ to find the threshold for giant components.

$$g(x) = \sum_{k=1}^{\infty} ka_k x^k$$

Now, $g(1) = \sum_{k=0}^{\infty} ka_k$ is the probability that a randomly chosen vertex is in a finite sized component. For $\delta = 0$, this is one, but for $\delta = 1$, the vertex created at time one is in a nonfinite size component. From the formula for the a_i's, we will derive the differential equation

$$g = -2\delta x g' + 2\delta x g g' + x$$

and then use the equation for g to determine the value $\delta_{critical}$ of δ at which the phase transition for the appearance of a nonfinite sized component occurs.

$$g' = \tfrac{1}{2\delta}\tfrac{1-\frac{g}{x}}{1-g}$$

Phase Transition for nonfinite components

A finite component is a component of size $1, 2, \ldots,$ which does not depend on t. If $g(1) = 1$ there are no infinite components. When $g(1) \neq 1$, then $1 - g(1)$ is the expected fraction of the vertices that are in nonfinite components. Potentially, there could be many such nonfinite components. But two fairly large components would merge into one. This leaves open the possibility of many components with some slowly growing function

of t.
To calculate $\delta_{critical}$ at $x = 1$, $g(1)$ is simply given by

$$g'(1) = \frac{1}{2\delta} \tag{1}$$

The quadratic $(g'(1))^2 - \frac{1}{2\delta}g'(1) + \frac{1}{2\delta}g(1) = 0$ has solutions

$$g'(1) = \frac{\frac{1}{2\delta} \pm \sqrt{\frac{1}{4\delta^2} - \frac{4}{2\delta}}}{2} = \frac{1 \pm \sqrt{1 - 8\delta}}{4\delta}. \tag{2}$$

We calculate all correct solutions for $0 \leq \delta \leq 1/8$ and $\delta > 1/8$ and conclude that (2) with the minus sign is the correct solution for $0 \leq \delta < 1/8$ and hence the critical value of $\delta$for the phase transition is $1/8$. As we shall see, this is different from the static case.
As the value of δ increased, the average size of the finite components increase from one to two when δ reaches the critical value of $1/8$. At $\delta = 1/8$, theaverage size of the finite components jumps to 4 and then decreases.

Comparison to static random graphs

In a static random graph with the same degree distribution as the graph in the growth model occurs a phase transition for $\delta = 1/4$.

3.2 Growth Model With Preferential Attachment

We take a model which generates a tree with power law distribution. Let $d_i(t)$ be the expected degree of the i^{th} vertex at time t. The sum of expected degrees of all vertices at time t is $2\delta t$. The degree of vertex i is governed by $\frac{d_i(t)}{2t}$. We can show that $d_i(t) = \delta\sqrt{\frac{t}{i}}$. The probability density $p(d)$ is a power law distribution.

4 Small World Graphs

For modeling real world phenomena, we adopt a simple model of a directed graph having local and long-distance edges. The probability that the long-distance edge from vertex u terminates at v is proportional to $1/d^r(u, v)$, $\quad d(u, v)$ is the distance from u to v, r is a constant.

$r > 2$ The legths of long-distance edges tend to be short so the probability of encountering a sufficiently long, long-distance edge is too low.

$r = 2$ Selecting the edge with end point closest to the destination finds a short path.

$r < 2$ The ends of long-distance edges tend to be uniformly distributed. Short paths exist but a polylog legth path is unlikely to encounter a long-distance edge whose end point is close to the destination.

Let $c_r(u) = \sum_v d^{-r}(u, v)$ be the normalizing constant.

For $r > 2, c_r(u)$ is lower bounded by 1

For $r = 2, c_r(u)$ is upper bounded by $\theta(\ln n)$

For $r < 2, c_r(u)$ is lower bounded by $\sum_{k=n/4}^{n/2} k^{1-r}$