Emo Welzl, Mittagsseminar, 15. Juni 1993

Random Independent Sets.

Given a graph $G = (V, E)$, $V = \{1, 2, \ldots, n\}$ and $E \subseteq \binom{V}{2}$, an independent set is a subset I of V with $E \cap \binom{I}{2} = \emptyset$. We all know that finding a maximum cardinality independent set is NP-hard, and actually there is no polynomial algorithm which approximates the maximum independent set up to a factor of $2^{O(\sqrt{\log n})}$, unless P=NP [AS1].

Here we consider two ways of taking random independent sets, and give bounds on how large they are on the average.

Random ordering. Let π be an ordering of the vertices of G. We define a set $S_\pi \subset V$, by letting $i \in S_\pi$ if and only if none of its neighbors precedes i in the order π. Clearly, S_π is an independent set[1] in G. What is its expected size, if π is chosen at random uniformly from all $n!$ permutation? For $i \in V$, let X_i be the random indicator variable for the event that i is in S_π. If d_i denotes the number of neighbors of i in G, then the probability of i to lie in S_π is $1/(d_i + 1)$. It follows that

$$\mathrm{E}(|S_\pi|) = \mathrm{E}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{E}(X_i) = \sum_{i=1}^{n} \frac{1}{d_i + 1} \ . \tag{1}$$

This bound is probably not very instructive, so let us try to conclude a bound in terms of $m = |E|$. To this end, observe that[2]

$$\left(\frac{\sum_{i=1}^{n} (d_i + 1)^{-1}}{n} \right)^{-1} \leq \frac{\sum_{i=1}^{n} d_i + 1}{n} = \frac{2m + n}{n} \ ,$$

and so

$$\mathrm{E}(|S_\pi|) \geq \frac{n^2}{2m + n} \ .$$

If G consists of k cliques of size s, then $m = k\binom{s}{2}$, and $n = ks$. Hence, the bound gives

$$\frac{k^2 s^2}{ks(s-1) + ks} = k \ ,$$

which is obviously tight. (Indeed, bound (1) implies already Turán's Theorem, see [AS1, page 81].) Note, that in an algorithmic setting, we could look at the vertices of G in the order given by π and add a vertex to a set T, if none of its neighbors are in T yet[3]. In this way we get a set $T_\pi \supseteq S_\pi$, which may be much larger than S_π. For example, if G is a complete bipartite graph with independent sets of size s, $n = 2s$, then $\mathrm{E}(|S_\pi|) = n/(s+1) = 2n/(n+2) < 2$, while T_π will always contain $n/2$ vertices.

Taking random subsets ... How many vertices can we choose, so that we can expect less than one edge among those vertices? If we choose a set R of r vertices at random (every set in $\binom{V}{r}$ with the same probability), then the probability for an edge $\{i, j\}$ to have both endpoints in R is

$$\frac{\binom{n-2}{r-2}}{\binom{n}{r}} = \frac{r(r-1)}{n(n-1)} \ ;$$

i.e.

$$\mathrm{E}(|\binom{R}{2} \cap E|) = \frac{mr(r-1)}{n(n-1)} \ .$$

If we want this quantity to be less than one, then we have to choose r to be roughly $n/\sqrt{m}$.

[1] If $\{i, j\}$ is an edge, and say $\pi(i) < \pi(j)$, then j is not in S_π.

[2] Harmonic mean cannot exceed arithmetic mean.

[3] This algorithm is due to P. K. Greedy.

. . . and remove some vertices again. Here is another approach. Let us fix a real number p, $0 \le p \le 1$. We choose a subset S by taking every vertex into S with probability[4] p. For each edge in $F = \binom{S}{2} \cap E$, we remove one endpoint from S and we obtain an independent set of size at least $|S| - |F|$ in this way. Let X be the random variable for $|S|$ and let Y be the random variable for $|F|$. We obtain

$$\mathrm{E}(X - Y) = \mathrm{E}(X) - \mathrm{E}(Y) = np - mp^2 \ ,$$

and this expression maximizes for $p = n/(2m)$. This yields $\mathrm{E}(X - Y) = n^2/(4m)$, which gets us reasonably close to the bound we obtained before. In particular, it is significantly better than the previous approach without removal (a trick due to Erdős).

There is no reason to choose the probabilities to be the same for all vertices. This eventually leads us to the following

Lemma 1 *The solution to the quadratic program*

$$\begin{array}{c} \textit{maximize } \sum_{i=1}^n p_i - \sum_{\{i,j\}\in E} p_i p_j \\ \textit{subject to } 0 \le p_i \le 1 \textit{ for all } 1 \le i \le n \end{array}$$

equals the maximal cardinality of an independent set.

First, if I is an independent set, then we can set $p_i = 1$ for $i \in I$ and $p_i = 0$ for $i \notin I$. Then the value of the program is exactly $|I|$. This shows that the optimal value of the program is at least the maximal cardinality of an independent set.

On the other hand, given some values of p_i in the allowed range $(0 \le p_i \le 1)$, then we choose a set S of vertices, vertex i with probability p_i. Then $\sum_{i=1}^n p_i$ is the expected number of vertices, and $\sum_{\{i,j\}\in E} p_i p_j$ is the expected number of edges with both endpoints in S. By the removal process suggested above, this gives an independent set whose expected size is at least the value of the program.

So what would be good values for the p_i's? It is clear, that we should always take vertices with degree 0, we cannot loose by taking vertices of degree 1, and in general, vertices of small degree seem to be preferable. So let us assume that there are no vertices of degree 0, which allows us to set $p_i = 1/d_i$. Observe that[5]

$$\sum_{\{i,j\}\in E} \frac{1}{d_i d_j} \le \frac{1}{2} \sum_{i=1}^n \left(\frac{1}{d_i}\right)^2 d_i = \frac{1}{2} \sum_{i=1}^n \frac{1}{d_i} \ ,$$

i.e.

$$\sum_{i=1}^n \frac{1}{d_i} - \sum_{\{i,j\}\in E} \frac{1}{d_i d_j} \ge \frac{1}{2} \sum_{i=1}^n \frac{1}{d_i} \ge \frac{n^2}{4m} \ .$$

In other words, we cannot lose by choosing the probabilities $p_i = 1/d_i$, and we will gain, unless all the d_i's are the same. ◻

[AS1] Noga Alon, Joel H. Spencer, *The Probabilistic Method*, John Wiley & Sons, Inc., New York (1992).
[AS2] Sanjeev Arora, Shmuel Safra, Probabilistic checking of proofs, a new characterization of NP, *in* "Proc. 33rd Annual Symp. on Foundations of Computer Science," (1992) 2–13

[4] This is similar to taking a random subset of size pn.
[5] Use Cauchy's inequality: $(\sum_{i=1}^n a_i b_i)^2 \le \sum_{i=1}^n a_i^2 \sum_{i=1}^n b_i^2$.