SUBLINEAR TIME AVERAGE DEGREE

Given $G = (V, E)$, where $|V| = n$, and connected can we estimate the average degree in G by inspecting less than all of the vertices?

Theorem. *Let d be the average degree in G and d_S be the average degree of a set S of s vertices chosen uniformly[1] from V. Then if $s \geq \beta\sqrt{n}/\varepsilon^{O(1)}$, $\varepsilon^{O(1)}$ simply means we can choose any constant that makes the proof workout, for some β to be fixed later*

$$P\left(d_S \geq \left(\frac{1}{2} - \varepsilon\right) \cdot d\right) \geq 1 - \frac{\varepsilon}{64} \tag{0.1}$$

For the moment assume the theorem and note that by Markov's inequality it's immediate that

$$P\left(d_S \leq (1+\varepsilon) \cdot d\right) \geq 1 - \frac{1}{1+\varepsilon} \geq \frac{\varepsilon}{2} \tag{0.2}$$

Why? Markov's inequality says

$$P\left(X \geq a\right) \leq \frac{E\left(X\right)}{a}$$

and by complement

$$1 - P\left(X < a\right) \leq \frac{E\left(X\right)}{a} \iff P\left(X < a\right) \geq 1 - \frac{E\left(X\right)}{a}$$

Then $X = d_S$ and $E\left(X\right) = d$ and $a = (1+\varepsilon) \cdot d$ gives

$$P\left(d_S < (1+\varepsilon) \cdot d\right) \geq 1 - \frac{d}{(1+\varepsilon) \cdot d} = 1 - \frac{1}{1+\varepsilon} = \frac{\varepsilon}{1+\varepsilon} \geq \frac{\varepsilon}{2}$$

What's the point? We can exploit these two inequalities to construct a $(2+\varepsilon)$-approximation algorithm[2]: pick $k = \frac{8}{\varepsilon}$ sets S_i of size s and output the set with smallest average degree. Using eqn. 0.2 (why eqn. 0.2 and not more obviously eqn. 0.1? To get the direction of the inequalitiy to work) the probability that all of the sets S_i overestimate average degree is

$$\begin{aligned}
\left(P\left(d_S \geq (1+\varepsilon) \cdot d\right)\right)^k &= \left(1 - P\left(d_S < (1+\varepsilon) \cdot d\right)\right)^k \\
&= \left(1 - \frac{\varepsilon}{2}\right)^k = \left(1 - \frac{\varepsilon}{2}\right)^{8/\varepsilon} \\
&= \left(\left(1 - \frac{1}{2\left(\frac{1}{\varepsilon}\right)}\right)^{2\left(\frac{1}{\varepsilon}\right)}\right)^4 \\
&\leq \left(e^{-1}\right)^4 \\
&\leq \frac{1}{8}
\end{aligned}$$

The probabilitiy that any one of the k sets underestimates the average degree is

$$\begin{aligned}
k \cdot P\left(d_S < \left(\frac{1}{2} - \varepsilon\right) \cdot d\right) &= k \cdot \left(1 - P\left(d_S \geq \left(\frac{1}{2} - \varepsilon\right) \cdot d\right)\right) \\
&\leq k \cdot \left(1 - \left(1 - \frac{\varepsilon}{64}\right)\right) \\
&\leq k \cdot \frac{\varepsilon}{64} = \frac{8}{\varepsilon} \cdot \frac{\varepsilon}{64} = \frac{1}{8}
\end{aligned}$$

[1] d_S is hence a random variable.

[2] $d_S \leq d \leq d_S\left(2+\varepsilon\right)$.

Hence with probability $1 - \frac{1}{8} - \frac{1}{8} = \frac{3}{4}$ both eqn. 0.1 and eqn. 0.2 will be satisfied. Finally $\varepsilon \to \varepsilon/2$ yields a $(2+\varepsilon)$-approximation algorithm.

Proof. Let H be the set of vertices with highest degree in G and $L = V \backslash H$. Fix ε such that $|H| = \sqrt{\varepsilon n}$. Let v_d be the degree of a vertex and firstly

$$\sum_{v \in L} v_d \geq \left(\frac{1}{2} - \varepsilon\right) \sum_{v \in V} v_d \tag{0.3}$$

i.e. $\sum_{v \in L} v_d$ is at least some non-vanishing proportion of total degree of G. Why is this the case? Edges incident on a vertex in L could contribute 2 to the sum: if $\{u, v\} \subset L$ then (u, v) is counted in u_d and v_d. But if either u or v is in H then (u, v) will be counted in either u_d or v_d but not both. This isn't so bad because it means each (u, v), as counted by $\sum_{v \in L} v_d$, is at least half as much as counted by $\sum_{v \in V} v_d$. If that were the whole story you'd have immediately

$$\sum_{v \in L} v_d \geq \frac{1}{2} \sum_{v \in V} v_d$$

Unfortunately edges (u, v) for which $\{u, v\} \subset H$ might be a supermajority of the edges in E (thereby making $\sum_{v \in L} v_d$ small again). But this can't happen because of how we chose ε: since $|H| = \sqrt{\varepsilon n}$ there can only be $\left(\sqrt{\varepsilon n}\right)^2 = \varepsilon n$ such edges[3], which is small relative $|E|$[4]. So

$$\sum_{v \in L} v_d \geq \frac{1}{2} \sum_{v \in V} v_d - 2\varepsilon n \geq \frac{1}{2} \sum_{v \in V} v_d - \varepsilon \sum_{v \in V} v_d \geq \left(\frac{1}{2} - \varepsilon\right) \sum_{v \in V} v_d$$

Full disclosure: I'm not certain here. Okay don't forget the goal here is to produce a lower bound on the average degree of the vertices that end up in our sample. Therefore without loss of generality we can assume all vertices are sampled[5] from L. Let S be that sample (with $|S| = s$) and let X_i be degree of the ith vertex[6] in S. With d_H being the vertex of lowest degree in H we have that for all $v \in L$ it's the case that $1 \leq v_d \leq d_H$ (since by definition all vertices in H have higher degree than all vertices in L) and hence $1 \leq X_i \leq d_H$. Then (*supposedly*) by Hoeffding's inequality

$$\begin{aligned} P\left(\sum_{i=1}^{s} X_i \leq (1-\varepsilon)\, E\left(\sum_{i=1}^{s} X_i\right)\right) &= P\left(\sum_{i=1}^{s} X_i - E\left(\sum_{i=1}^{s} X_i\right) \leq -\varepsilon E\left(\sum_{i=1}^{s} X_i\right)\right) \\ &\leq \exp\left(-\frac{2\varepsilon^2 \left(E\left(\sum_{i=1}^{s} X_i\right)\right)^2}{\sum_{i=1}^{s} \left(d_H - 1\right)^2}\right) \\ &\cdots \\ &\exp\left(-\frac{\varepsilon^2 \left(E\left(\sum_{i=1}^{s} X_i\right)\right)}{d_H}\right) \end{aligned} \tag{0.4}$$

I have no idea how they got the right hand side. For Bernoulli random X_i you use the Angluin/Valiant[7] inequality to get

$$P\left(\sum_{i=1}^{s} X_i \leq (1-\varepsilon)\, E\left(\sum_{i=1}^{s} X_i\right)\right) \leq \exp\left(-\frac{\varepsilon^2 E\left(\sum_{i=1}^{s} X_i\right)}{2}\right)$$

---

[3] If H is completely connected, i.e. a clique, then every vertex in H is connected to every other, i.e. $\sqrt{\varepsilon n}$ edges for everyone of the $\sqrt{\varepsilon n}$ vertices.

[4] At least $n - 1$ since the graph is connected (a minimum spanning tree on n vertices has $n - 1$ edges)

[5] Since all vertices in H have higher degree, any lower bound produced by sampling L will also be a lower bound for average degree of vertices in H.

[6] A random variable since S is a sample.

[7] D. Angluin and L.G. Valiant, "Fast probabilistic algorithms for Hamiltonian circuits and matchings", *Journal of Computer and System Sciences*, No. 19, 1979, pp. 155–193.

But that's still not correct because it's missing d_H. Anyway moving on. This has to be interpreted now in terms of d_S and d. Firstly $d_S = \sum_{i=1}^{s} X_i$ and $d = E(d_S) = E\left(\sum_{i=1}^{s} X_i\right)$. So

$$P(d_S \leq (1-\varepsilon)d) \leq 1 - \exp\left(-\frac{\varepsilon^2 E\left(\sum_{i=1}^{s} X_i\right)}{d_H}\right)$$

Since any vertex in H has degree at least d_H the average degree has to be at least[8] $d_H \cdot \left(\frac{|H|}{n}\right)$. Hence by eqn. 0.3 we have that

$$E(X_i) \approx \frac{1}{|S|}\sum_{v\in S} v_d \geq \frac{1}{|L|}\sum_{v\in L} v_d \geq \frac{1}{|L|}\left(\frac{1}{2}-\varepsilon\right)\sum_{v\in V} v_d \geq \frac{1}{|V|}\left(\frac{1}{2}-\varepsilon\right)\sum_{v\in V} v_d \geq \left(\frac{1}{2}-\varepsilon\right) d_H \left(\frac{|H|}{n}\right)$$

The $\approx$ comes from the law of large numbers [9] I think and the first inequality is because of ???. By linearity of expectation we have that

$$E(\Sigma_{i=1}^{s} X_i) \geq s\left(\frac{1}{2}-\varepsilon\right) d_H \left(\frac{|H|}{n}\right)$$

Given our choice of $s \geq \beta\sqrt{n}/\varepsilon^{O(1)}$

$$\begin{aligned}
E(\Sigma_{i=1}^{s} X_i) &\geq s\left(\frac{1}{2}-\varepsilon\right) d_H \left(\frac{|H|}{n}\right) \\
&\geq \left(\frac{1}{2}-\varepsilon\right) d_H \left(\frac{|H|}{n}\right) \frac{\beta\sqrt{n}}{\varepsilon^{O(1)}} \\
&= \left(\frac{1}{2}-\varepsilon\right) d_H \left(\frac{|H|}{\sqrt{n}\varepsilon^{O(1)}}\right)\beta \\
&= \left(\frac{1}{2}-\varepsilon\right) d_H \left(\frac{|H|}{\sqrt{n}\varepsilon\varepsilon^{O(1)}}\right)\beta \\
&= \left(\frac{1}{2}-\varepsilon\right)\left(\frac{d_H\beta}{\varepsilon^{O(1)}}\right)
\end{aligned}$$

Note between lines 4 and 5 in the above the constant in $\varepsilon^{O(1)}$ changes. Therefore

$$\begin{aligned}
P(d_S \leq (1-\varepsilon)d) &\leq \exp\left(-\frac{\varepsilon^2 E\left(\sum_{i=1}^{s} X_i\right)}{d_H}\right) \\
&\leq \exp\left(-\frac{\varepsilon^2\left(\frac{1}{2}-\varepsilon\right)\left(\frac{d_H\beta}{\varepsilon^{O(1)}}\right)}{d_H}\right) \\
&\leq \exp\left(-\beta\left(\frac{1}{2}-\varepsilon\right)\right)
\end{aligned}$$

Choose β such that

$$\exp\left(-\beta\left(\frac{1}{2}-\varepsilon\right)\right) \leq \frac{\varepsilon}{64}$$

and taking the compliment we get

$$P(d_S \geq (1-\varepsilon)d) \geq 1 - \frac{\varepsilon}{64}$$

Finally since $B \Leftarrow A$ implies $P(B) \geq P(A)$ we have that

$$P\left(d_S \geq \left(\frac{1}{2}-\varepsilon\right)d\right) \geq P(d_S \geq (1-\varepsilon)d) \geq 1 - \frac{\varepsilon}{64}$$

□

[8] $d = \frac{1}{n}\sum_{i=1}^{n} v_{d,i} = \frac{1}{n}\sum_{v\in L} v_d + \frac{1}{n}\sum_{v\in H} v_d \geq \frac{1}{n}\sum_{v\in L} v_d + \frac{1}{n}|H|\,d_H$

[9] The law of large numbers: for all ϵ and random variables X_i $\lim_{n\to\infty}\left(\left|\bar{X}_n - E\left(\bar{X}_n\right)\right| > \varepsilon\right) = 0$