Estimating the Average Degree in an Undirected Graph[1]

- *Sublinear Graph Algorithms.* In this lecture, we see an algorithm to estimate a statistic about graphs making *sub-linear* number of "accesses" to it. More precisely, there is an unknown graph $G = (V, E)$ on $n = |V|$ vertices, and we wish to determine the *average degree* $d_{\mathsf{avg}} := \frac{2m}{n}$ of the graph where m is the number of edges in E, in $o(n)$ time. Note that this rules out the trivial algorithm of querying every vertex and obtaining their degrees. Before we go into the algorithm, we need to precisely state how we are allowed to access the graph.

- *The Graph Query Model.* We assume a *query access model* for the graph: imagine this huge graph G is *not* sitting in the RAM of our computer, but rather is owned by some third party allowing us the following three APIs to access it. This model was formalized[2] by Oded Goldreich and Dana Ron.

  G1. *Random Access.* We are allowed to access a random vertex v uniformly at random from V.

  G2. *Degree Queries.*We are allowed to query the *degree* $\deg(v)$ of any vertex v.

  G3. *Neighbor Queries.* For any vertex v, and for any integer $1 \leq i \leq \deg(v)$, we can ask for the ith neighbor of v. The order is an arbitrary but fixed order.

- *A First Try.* The first algorithm that probably comes to everyone's mind is the following : sample a random vertex v (using G1) and query its degree $\deg(v)$ (using G2). Crucially note that we haven't used the power of (G3) at all. Set $\widehat{\mathsf{est}} := \deg(v)$. Clearly,

$$\mathbf{Exp}[\widehat{\mathsf{est}}] = \sum_{v \in V} \deg(v) \cdot \frac{1}{n} = d_{\mathsf{avg}}$$

  Now, let's look at the *variance* $\mathbf{Var}[\widehat{\mathsf{est}}] = \mathbf{Exp}[\widehat{\mathsf{est}}^2] - \mathbf{Exp}^2[\widehat{\mathsf{est}}]$. We see that

$$\mathbf{Var}[\widehat{\mathsf{est}}] \ \leq \mathbf{Exp}[\widehat{\mathsf{est}}^2] = \sum_{v \in V} \deg^2(v) \cdot \frac{1}{n} \leq \ \Delta \cdot d_{\mathsf{avg}}$$

  where Δ is the *maximum* degree in G. Therefore, if we let $\rho := \frac{\Delta}{d_{\mathsf{avg}}}$, then we get $\frac{\mathbf{Var}[\widehat{\mathsf{est}}]}{\mathbf{Exp}^2[\widehat{\mathsf{est}}]} \leq \rho$, which in turn implies $O(\frac{\rho \cdot \ln(1/\delta)}{\varepsilon^2})$ samples would lead to an (ε, δ)-multiplicative approximation.

  The issue is that ρ of a graph can be really large, and indeed, it is really large for many real-world graphs such as the web graph. The paradigmatic bad example is the *star graph* which has one center vertex of degree n and n other leaf vertices of degree 1. Here, $\rho = n$, but $d_{\mathsf{avg}} \approx 2$. Thus, $\rho = \frac{n}{2}$, which gives a useless bound above.

  In fact the star graph tells us more. It shows that *any* algorithm which makes $o(n)$ queries, will almost surely *not* query the center node, and thus will only see $\deg(v) = 1$. The reasonable estimate here would be to set $\widehat{\mathsf{est}} = 1$, and this will be $\approx 50\%$ off. Indeed, this can be formalized into the following theorem.

[1] *Lecture notes by Deeparnab Chakrabarty. Last modified : 15th April, 2021*
These have not gone through scrutiny and may contain errors. If you find any, or have any other comments, please email me at deeparnab@dartmouth.edu. Highly appreciated!

[2] *Approximating Average Parameters of Graphs*, Random Structures & Algorithms, 32: 473–493, 2008

Theorem 1 (Feige[3])**.** Any algorithm with only (G1) and (G2) queries making an $(2-c)$-approximation to the average degree, for any constant c, must make $\Omega(n)$ queries.

In his paper, Feige in fact proved that for any graph $\approx O(\sqrt{n})$ samples of $\widehat{\mathsf{est}}$ *can* be used to give a $(2+\varepsilon)$-approximation to d_{avg}. I may form a problem in the problem set exploring this. But today's lecture is about getting an (ε, δ)-multiplicative approximation via a very simple, but clever, algorithm which uses the power of neighbor queries(G3). This result was first obtained by Goldreich and Ron in their paper where they formalized the query model. This simpler analysis is essentially due[4] to Talya Eden, Dana Ron, and C. Seshadhri, although they look at a much more involved problem (which is a great reading project).

- *The Cute Idea.* We know when $\widehat{\mathsf{est}}$ is bad: it is when $\rho := \Delta/d_{\mathsf{avg}}$ is high. Like in the star graph. The presence of high-degree vertices leads to high variance. The main difference in the algorithm is whenever one encounters a vertex v, then instead of setting the estimate to $\deg(v)$ right away, one first queries a random neighbor x of v and sets the estimate to $2\deg(v)$ iff $\deg(x) > \deg(v)$. In other words, if the degree of v is so big that most of its neighbors have small degree, it ignores that vertex. This way it can control the variance. And the factor 2 is precisely what keeps the estimate unbiased because the number of edges can be counted by only considering the degree of the "lower degree" endpoint. Let's be precise.

  There is one annoying technicality we need to take care of regarding tie-breaking. In the first read, assume all degrees are distinct and move on. Otherwise, we assume there is a function id which assigns each vertex a unique id in a ordered set. If two vertices have the same degree, then we use the id to break ties.

```
1:  procedure DegreeEstimate(G):
2:      Sample a random vertex v in G via (G1).
3:      Query the degree deg(v) via (G2).
4:      Randomly sample i ∈ {1, 2, ..., deg(v)} and query the ith neighbor x of v via (G3).
5:      if deg(x) ≻ deg(v) then:
6:          ▷ We say deg(x) ≻ deg(v) if deg(x) > deg(v) OR if deg(x) = deg(v) but id(x) >
    id(v).
7:          Set est = 2 deg(v).
8:      else:
9:          Set est = 0.
10:     return est.
```

- *Analysis.* To analyze the algorithm, for every vertex v, let $\deg^+(v) := |\{x \sim v : \deg(x) \succ \deg(v)\}|$ count the number of neighbors of v with higher degree. We establish two simple but key claims about $\deg^+(v)$: one takes care of the expectation, and the other takes care of the variance.

  Claim 1. $\sum_{v \in V} \deg^+(v) = m$

[3] *On Sums of Independent Random Variables with Unbounded Variance, and Estimating the Average Degree of an Unknown Graph*, U. Feige, SIAM J. Comp., 35(4): 964–984

[4] *Sublinear Time Estimation of Degree Distribution Moments: The Degeneracy Connection*, T. Eden, D. Ron, C. Seshadhri. SIAM J. Discret. Math. 33(4): 2267-2285 (2019)

Proof. For every edge (x, y) *direct* it from the vertex x with $\deg(x) \succ \deg(y)$ to the vertex y. In this directed graph, the out-degree of every vertex v is precisely $\deg^+(v)$. The claim follows from the handshake lemma. □

Claim 2. For any vertex v, $\deg^+(v) \leq \sqrt{2m}$.

Proof. Suppose $M = \deg^+(v)$ is $> \sqrt{2m}$. Let $x_1, \ldots, x_M$ be the $M > \sqrt{2m}$ neighbors of v with $\deg(x_i) \succ \deg(v)$. We get $\sum_{i=1}^{M} \deg(x) \geq M \deg(v) \geq M \deg^+(v) \geq M^2 > 2m$. This contradicts that the sum of degrees of all vertices is $= 2m$. □

Now, notice

$$\mathbf{Exp}[\widehat{\mathsf{est}}] = \sum_{v \in V} 2\deg(v) \cdot \frac{1}{n} \cdot \mathop{\mathbf{Pr}}_{x \sim v}[\deg(x) \succ \deg(v)] = \sum_{v \in V} \frac{2\deg(v)}{n} \cdot \frac{\deg^+(v)}{\deg(v)} \underbrace{=}_{\text{Claim 1}} \frac{2m}{n} = d_{\mathsf{avg}}$$

And,

$$\mathbf{Exp}[\widehat{\mathsf{est}}^2] = \sum_{v \in V} \frac{4\deg^2(v)}{n} \cdot \frac{\deg^+(v)}{\deg(v)} = \frac{4}{n} \sum_{v \in V} \deg(v) \deg^+(v) \underbrace{\leq}_{\text{Claim 2}} \frac{8m\sqrt{2m}}{n}$$

Using the fact that $d_{\mathsf{avg}} = \frac{2m}{n}$, we can rewrite this as

$$\mathbf{Var}[\widehat{\mathsf{est}}] \leq \mathbf{Exp}[\widehat{\mathsf{est}}^2] \leq \frac{8m\sqrt{2m}}{n} = \frac{2\sqrt{2}n}{\sqrt{m}} \cdot \left(\frac{2m}{n}\right)^2 = \frac{2\sqrt{2}n}{\sqrt{m}} \cdot \left(\mathbf{Exp}[\widehat{\mathsf{est}}]\right)^2$$

Now suppose we knew a *lower bound* d_0 to the average degree. For instance, suppose we knew there were no isolated vertices, then $d_0 = 1$. Then, $m \geq nd_0$ giving us $\frac{n}{\sqrt{m}} \leq \sqrt{\frac{n}{d_0}}$. Thus, using the boosting theorem we get the desired theorem.

Theorem 2. Let d_0 be a known *lower bound* to the average degree. Using $O\left(\sqrt{\frac{n}{d_0}} \cdot \frac{\ln(1/\delta)}{\varepsilon^2}\right)$ samples of DEGREEESTIMATE, one can obtain an (ε, δ)-multiplicative estimate to the average degree d_{avg}.

Learning Tidbits:

- Algorithm Design: *If the first unbiased estimate has high-variance, then try to see where the high-variance arises from, and try to cull them out. This is a general and vague idea, and whether it works depends on the problem at hand. Here it worked because the number of edges (or twice the average degree) could be counted by summing up all degrees and dividing by 2, or only summing up degrees in such a way that for every edge only one endpoint is counted. And in this case, the lower degree is counted.*
- Analysis: *The idea that $\deg^+(v)$ can't be too large is indeed the key claim. It's simple, but key.*