Lecture 4:

Estimating the Average Degree

Course: Algorithms for Big Data

Instructor: Hossein Jowhari

Department of Computer Science and Statistics
Faculty of Mathematics
K. N. Toosi University of Technology

Spring 2021

Outline

- An application of Chernoff bound
- Estimating the average degree (continued)

Recap from previous lecture

Chernoff Bound: Let $0 \leq \epsilon \leq 1$. Suppose $Y_1, \ldots, Y_t$ are independent random variables taking values in the interval $[0, 1]$. Let $Y = \sum_{i=1}^{t} Y_i$. Then

$$Pr\Big(\, Y \; \leq (1 - \epsilon) E[Y] \Big) \leq e^{-\frac{\epsilon^2 E[Y]}{2}}$$

$$Pr\Big(\, Y \; \geq (1 + \epsilon) E[Y] \Big) \leq e^{-\frac{\epsilon^2 E[Y]}{3}}$$

$$Pr\Big(\, |Y - E[Y]| \; \geq \epsilon E[Y] \Big) \leq 2e^{-\frac{\epsilon^2 E[Y]}{3}}$$

An application of Chernoff bound

Amplifying the success probability

Suppose we have a randomized algorithm A that processes the input data D and approximate some $f(D)$ where

$$|A(D) - f(D)| \leq \epsilon f(D) \text{ with probability at least } 3/4.$$

How to amplify the success probability of A?

We want to have a randomized algorithm A' with error probability $\delta << 1/4$.

Idea: Run A on input data D, $t = \Omega(\ln(\frac{1}{\delta}))$ times and output the median of the outcomes.

The median of n elements is the element with rank $\lceil \frac{n}{2} \rceil$. The rank of an element is its position in the sorted list.

Example: $\mathrm{median}(5, 2, 14, 21, 15, 10, 6) = 10$

Suppose a_i is the outcome of i-th run of algorithm A.

Final Ouput: $\mathrm{median}(a_1, a_2, \ldots, a_t)$.

Each (independent) run of A succeeds with probability $3/4$. Therefore we have

$$Pr(|a_i - f(D)| \geq \epsilon f(A)) \leq 1/4.$$

We define $X_i = 1$ if i-th run is good (its error is less than $\epsilon f(A)$), otherwise we let $X_i = 0$.

$X = X_1 + \ldots + X_t$ is the number of good outcomes in t repetitions.

We define the events:

$$
\begin{aligned}
E_1 &= \text{The median of } \{a_1, \ldots, a_t\} \text{ is bad} \\
E_2 &= \text{Less than } t/2 \text{ repetitions are good} \quad \Rightarrow \; X < t/2
\end{aligned}
$$

Important Observation: If the event E_1 happens then the event E_2 has happened. Therefore $Pr(E_1) \leq Pr(E_2)$.

$$\text{Median is too small: } \underbrace{5, 5, 8, 8, 11, \ldots, 78}_{\text{bad}}, \underbrace{\color{red}{79}}_{\text{bad}}, 83, \ldots, 121, 124, 130$$

$$\text{Median is too big: } 5, 5, 8, 8, 11, \ldots, 78, \underbrace{\color{red}{79}}_{\text{bad}}, \underbrace{83, \ldots, 121, 124, 130}_{\text{bad}}$$

It follows, $Pr(\text{median is bad}) \leq Pr(X < t/2)$.

Since $E[X] \geq \frac{3}{4}t$, using Chernoff bound, we get

$$Pr(X < t/2) \leq Pr(X < (1 - \frac{1}{3})E[X]) \leq e^{-(\frac{1}{3})^2(\frac{3}{4}t)(\frac{1}{2})} \leq \delta$$

Finally we get

$$t \geq \frac{72}{3}\ln(\frac{1}{\delta})$$

Estimating the average degree (continued)

A slight change in the algorithm

The value of repetition

Basic estimator: Compute

$$X = \frac{1}{s}(X_1 + \ldots + X_s)$$

where each X_i is the degree of a random vertex.

The New estimator: Repeat the Basic estimator $\frac{8}{\epsilon}$ times and output the smallest outcome.

$$\overbrace{S_1, \quad S_2, \quad \ldots, \quad S_{8/\epsilon}}^{\text{samples in each repetition}}$$

$$S_i \subseteq V, \qquad\qquad |S_i| = s$$

Consider a repetition of the basic estimator. By Markov inequality, we have

$$Pr\big(X \geq (1+\epsilon)E[X]\big) \leq \frac{1}{1+\epsilon} \leq 1 - \frac{\epsilon}{2}$$

What is the probability that in all repetitions, the basic estimator X is larger than $(1+\epsilon)E[X]$?

$$\leq (1 - \frac{\epsilon}{2})^{8/\epsilon} \leq \frac{1}{8}, \qquad (1 - \frac{1}{n})^{n} \leq 1 - 1/e$$

Therefore with probability $1 - \frac{1}{8}$, the smallest outcome is not bigger than $(1+\epsilon)E[X]$. It does not overestimate by large.

The probability of underestimation: Now we need to bound the probability that the outcome of a repetition falls below $\alpha E[X]$ for $\alpha < 1$.

Suppose we are able to prove the probability that the outcome of a single repetition falls below $\alpha E[X]$ (bad event) is at most $\frac{\epsilon}{64}$.

The probability that at least one repetition is bad is at most $\frac{8}{\epsilon} \times \frac{\epsilon}{64} = \frac{1}{8}$. (Union bound)

Therefore the probability of bad events happening (overestimation and underestimation) is at most $\frac{1}{8} + \frac{1}{8} = \frac{1}{4}$.

Main Idea: Recall the result we got from Chernoff bound:

$$\text{number of samples} = s \geq \frac{3}{\epsilon^2}\ln(\frac{1}{2\delta})\frac{d_{max}}{d}$$

This say if the term $\frac{d_{max}}{d}$ is small then the number of required samples would be small.

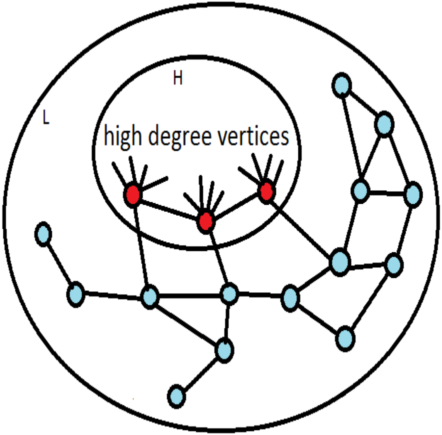
Suppose in the graph $G = (V, E)$, we show there is always a subset $L \subset V$ with the following properties:

1. The average degree in L is at least $(\frac{1}{2} - \epsilon)d$.
2. Every vertex outside L has degree $\geq d_{max}(L)$.
3. $\frac{d_{max}(L)}{d(L)} \approx O(\sqrt{n})$

$d(L)=$ average degree in L

$d_{max}(L) =$ maximum degree in L

Bounding the probability of underestimation:



Let H be the set of $\sqrt{\epsilon n}$ vertices with highest degrees.

Let L be the rest of vertices.

Let d_H be the smallest degree in H.

Note: total degree in graph is $2m$.

Claim: Average degree in L is at least $(\frac{1}{2} - \epsilon)d$

Proof: Lets calculate the total degree in L. In the process,

- Every edge with both endpoint in L is counted twice.
- Every edge with one endpoint in L and the other in H is counted once.
- The edges inside H are not counted.

How many edges can be inside H?

At most $\sqrt{\epsilon n} \times \sqrt{\epsilon n} = \epsilon n$ edges.

This is at most ϵm assuming $m \geq n$. Therefore total degree in L is at least $2m - \frac{1}{2}(2m) - 2\epsilon m = (\frac{1}{2} - \epsilon)2m$.

Since $|L| \leq n$, the average degree in L is at least

$$\frac{(\frac{1}{2} - \epsilon)2m}{|L|} \geq \frac{(\frac{1}{2} - \epsilon)2m}{n} = (\frac{1}{2} - \epsilon)d$$

□

Important Observation: Since we want to bound the probability of underestimation, we can safely assume that the estimator only samples the vertices in L.

(this is the worst-case scenario.)

Lets apply the Chernoff bound to the vertices in L.

$E \rightarrow$ the event that we sample only vertices in L

$$E[X \mid E] = d(L)$$

Note that $d_{max}(L) \leq d_H$. Conditioned on E, we have

$$Pr\big(X \leq (1-\epsilon)L(d)\big) \leq e^{-\frac{\epsilon^2 s}{3}\frac{d(L)}{d_H}}$$

Since the vertices in H have degree at least d_H,

$$d \geq \frac{|H| d_H}{n}$$

We also know that

$$d(L) \ \geq \ (\frac{1}{2} - \epsilon) d \ \geq \ (\frac{1}{2} - \epsilon) \frac{|H| d_H}{n}$$

We want

$$e^{-\frac{\epsilon^2 s}{2} \frac{d(L)}{d_H}} \leq e^{-\frac{\epsilon^2 s}{2} \frac{(\frac{1}{2} - \epsilon) \frac{|H| d_H}{n}}{d_H}} = e^{-\frac{\epsilon^2}{2} (\frac{1}{2} - \epsilon) \frac{s|H|}{n}}$$

to be smaller than $e^{-5 - \ln \frac{1}{\epsilon}} \leq \frac{\epsilon}{64}$.

Let assume $\epsilon < \frac{1}{4}$. Setting $s \geq \frac{8n}{\epsilon^2 |H|} (5 + \ln \frac{1}{\epsilon}) = O(\ln(\frac{1}{\epsilon}) \epsilon^{-2.5} \sqrt{n})$ works. (Note $|H| = \sqrt{\epsilon n}$)

Summing up

Let X be the outcome of a single repetition. We showed if we set $s = \beta \ln(\frac{1}{\epsilon})\epsilon^{-2.5}\sqrt{n}$ when β is a large enough constant, we have

$$Pr\Big(X \leq (1-\epsilon)L(d)\Big) \leq \frac{\epsilon}{64}$$

Let X^* be the repetition with smallest outcome.

$$Pr\Big(X^* \leq (1-\epsilon)(\frac{1}{2}-\epsilon)d\Big) \;\leq\; Pr\Big(X^* \leq (1-\epsilon)L(d)\Big) \;\leq\; \frac{8}{\epsilon}\frac{\epsilon}{64} = \frac{1}{8}$$

We also showed that

$$Pr\Big(X^* \geq (1+\epsilon)d\Big) \leq \frac{1}{8}$$

$$Pr\Big((1-\epsilon)(\frac{1}{2}-\epsilon)d \le X^* \le (1+\epsilon)d\Big) \ge 1-\frac{1}{8}-\frac{1}{8}$$

$$Pr\Big((\frac{1}{2}-\frac{3}{2}\epsilon)d \ \le \ X^* \ \le \ (1+\epsilon)d\Big) \ge \frac{3}{4}$$

To eliminate underestimation, we can return $X' = \frac{2}{1-3\epsilon}X^*$ instead. We have

$$Pr\Big(d \ \le \ X' \ \le \ 2\frac{1+\epsilon}{1-3\epsilon}d\Big) \ge \frac{3}{4}$$

$$Pr\Big(d \ \le \ X' \ \le \ (2+O(\epsilon))d\Big) \ge \frac{3}{4}$$

Query complexity: Number of degree queries we made in total is bounded by

$$\frac{8}{\epsilon} \times s = \frac{8}{\epsilon}\beta\ln(\frac{1}{\epsilon})\epsilon^{-2.5}\sqrt{n} = O(\ln(\frac{1}{\epsilon})\epsilon^{-3.5}\sqrt{n})$$

Theorem: Let $0 < \epsilon < 1/4$. Given a graph $G = (V, E)$ where $|V| = n$ and $|E| \geq n$, there is an algorithm that makes $O(\ln(\frac{1}{\epsilon})\epsilon^{-3.5}\sqrt{n})$ random degree queries and with probability at least $3/4$ returns a $2 + O(\epsilon)$ factor approximation of the average degree of the graph.

Feige's result: Given a graph $G = (V, E)$ where $|V| = n$ and $|E| \geq n$, there is an algorithm that makes $O(\epsilon^{-1}\sqrt{n})$ random degree queries and with probability at least $3/4$ returns a $2 + O(\epsilon)$ factor approximation of the average degree of the graph.

Using different types of queries (for example neighbor queries) we can get more efficient algorithms.

See Goldreich and Ron's and Dasgupta, Kumar, Sarlos's paper in the suggested reading list.