COL751 - Lecture 2

1 Probability Theory (Puzzles)

Question 1 Given is an $(n+1)$ length stick with n joints. The stick is dropped from certain height, due to which each joint breaks with probability p independent of other joints. What is the expected number of pieces into which the stick breaks ?

Solution: Introduce an indicator variable X_i for each joint $i \in [1, n]$ as below:

$$X_i \;=\; \begin{cases} 1 & \text{joint } i \text{ breaks,} \\ 0 & \text{otherwise.} \end{cases}$$

Then the number of resultant pieces will be $1 + \sum_{i=1}^{n} X_i$.

So, by linearity of expectation, the expected number of pieces is $E(1 + \sum_{i=1}^{n} x_i) = 1 + \sum_{i=1}^{n} E(x_i) = 1 + np$. □

Question 2 Let G be a graph with $3n$ vertices having n disjoint paths of length two. We perform following experiment.
Repeat following until no degree one vertices remain: Pick two vertices randomly uniformly and connect them by adding an edge. What is the expected number of cycles at the end ?

Solution: Introduce an indicator variable X_i for each joint $i \in [1, n]$ as below:

$$X_i \;=\; \begin{cases} 1 & \text{in step } i \text{ cycle is created,} \\ 0 & \text{otherwise.} \end{cases}$$

Then the number of cycles created will be $Y := \sum_{i=1}^{n} X_i$.

1. Before beginning of step i, there are exactly $n-i+1$ vertex disjoint paths. This holds because in each step we either create a cycle, or merge two paths.

2. If there were k vertex disjoint paths at a given stage, then the probability of creating a cycle will be
$$\frac{k}{{}^{2k}C_2} \;=\; \frac{1}{2k-1} \;=\; \frac{1/2}{(k-1/2)}.$$

3. So the expected number of cycles is:
$$E(Y) \;=\; \sum_{i=1}^{n} E(X_i) \;=\; \sum_{i=1}^{n} \frac{\frac{1}{2}}{n-i+\frac{1}{2}} \;\approx\; H_n/2.$$

Thus, the expected number of cycles will be $\Theta(\log n)$. $\square$

Homework Can you prove that the number of cycles created are bounded by $O(\log^2 n)$ with probability at least $1 - 1/n$?

2 Covering high degree vertices

In this section we will consider a basic application of random sampling.

Let $G = (V, E)$ be an n vertex graph, and let R be a uniformly random set of size of size $\frac{4n \log n}{d}$. Observe that for each vertex v of degree at least d, the expected number of neighbors of v lying in the random set R is at least one. We will see in the following theorem that with good enough probability, the set R will contain at least one neighbor of each and every vertex in G with degree at least d .

Theorem 1 *Let $G = (V, E)$ be an n vertex graph, and let R be a uniformly random set of size of size $\frac{4n \log n}{d}$. Then, with probability at least $(1 - 1/n^3)$ the following holds:*

For every $v \in V$ with degree $\geqslant d$, the set R contains a neighbor of v.

Proof: Let $k = \frac{4n \ln n}{d}$ and V_{heavy} be set of all vertices of degree at least d.

We are picking a random set $R = \{r_1, \ldots, r_k\}$ of size k such that each r_i is a uniformly random vertex from V and is chosen independent of other vertices in R.

Introduce an event $\mathcal{E}_v$ for every vertex $v \in V_{heavy}$ as below:

$$\mathcal{E}_v \ := \ \text{The event that } R \text{ contains no neighbor of } v.$$

Further, let $\mathcal{E}$ be the event that there exists at least one vertex $v \in V_{heavy}$ such that R has zero neighbor of R.

Our aim is to show that probability of $\mathcal{E}$ is small.

1. We first observe $\mathcal{E} = \cup_{v \in V_{heavy}} \mathcal{E}_v$.
2. Next note that for a particular vertex $v \in V_{heavy}$, the probability that none of the neighbors of v lie in R is

$$\text{Prob}(\mathcal{E}_v) \ = \ \text{Prob}\big(r_1, \ldots, r_k \notin N(v)\big) \ = \ \prod_{i=1}^{k} \text{Prob}\big(r_i \notin N(v)\big).$$

3. For any $v \in V_{heavy}$ and any $r_i \in R$,

$$\text{Prob}\big(r_i \notin N(v)\big) \ = \ \Big(\frac{n - |N(v)|}{n}\Big)$$

4. The probability of $\mathcal{E}_v$, for $v \in V_{heavy}$, is upper bounded by

$$\prod_{i=1}^{k} \Big(\frac{n - |N(v)|}{n}\Big) \;\leqslant\; \Big(1 - \frac{d}{n}\Big)^{\frac{4n \ln n}{d}} \;\leqslant\; e^{-4 \ln n} \;=\; \frac{1}{n^4}.$$

5. By union bound, we have:

$$\text{Prob}(\mathcal{E}) \;\leqslant\; \sum_{v \in V} \text{Prob}(\mathcal{E}_v) \;\leqslant\; \sum_{v \in V} \Big(\frac{1}{n^4}\Big) \;\leqslant\; \frac{1}{n^3}.$$

Therefore, with probability at least $(1 - 1/n^3)$ the set R has a non-empty intersection with neighborhood of all vertices in V_{heavy}. □

Remark A more general application of sampling in computing (random) hitting sets is given in Practice sheet 1. There is also an alternate deterministic construction for hitting set based on greedy approach that has $O(nd)$ running time, see here.

3 Subgraph preserving distances with +2 additive error

In this section we see an application of sampling in constructing sparse distance preserves that for any pair $(x, y) \in V \times V$ satisfy the following condition:

$$dist(x, y, G) \;\leqslant\; dist(x, y, H) \;\leqslant\; dist(x, y, G) + 2.$$

Let $G = (V, E)$ be an undirected connected graph with n vertices and m edge. The construction of sparse subgraph H is as follows:

1. Initialize $H = (V, E_H)$ as empty graph.
2. Partition V into two parts:
   V_{light}: All vertices with degree at most $\sqrt{n}$, and
   V_{heavy}: All vertices with degree more than $\sqrt{n}$.
3. For $x \in V_{light}$, add to H all edges incident to v in G.
4. Pick a set R of size $4\sqrt{n} \log n$ such that R has non-empty intersection with neighborhood of each vertex in V_{heavy}. (See Section 2).
5. For each $v \in R$, compute the BFS tree rooted at v in G, say T_v, and add the $n - 1$ edges of T_v to H.

Lemma 2 *The number of edges in H is $O(n\sqrt{n} \log n)$.*

Lemma 3 *For any $x, y \in V$, $dist(x, y, H) \leqslant dist(x, y, G) + 2$.*

Proof: Consider two vertices $x, y \in V$, and let P be any shortest path from x to y in G.

Case 1: All vertices of P lies in V_{light}. In this case all edges in P lie in H as well, and thus $dist(x, y, H) = dist(x, y, G)$.

Case 2: At least one vertex of P, say w, does not lie in V_{light}. In such a case $N(w) \cap R$ is non-empty, as $w \in V_{heavy}$.

Let r be a neighbor of w lying in set R. Then, by triangle inequality, we have

$$dist(x, y, H) \leqslant dist(x, r, H) + dist(r, y, H).$$

Since, the distance of r from x, y is identical in graphs G and H, we get:

$$\begin{aligned} dist(x, y, H) &\leqslant dist(x, r, H) + dist(r, y, H) \\ &\leqslant dist(x, r, G) + dist(r, y, G) \\ &\leqslant dist(x, w, G) + 1 + dist(w, y, G) + 1 \\ &= dist(x, y, G) + 2. \end{aligned}$$

This proves that distances are stretched by an additive factor of at most two. □

Due to Lemma 2 and Lemma 3, the following theorem is immediate.

Theorem 4 *For any unweighted undirected graph $G = (V, E)$ we can construct in $O(n^{2.5} \log n)$ time a subgraph graph H with $O(n^{1.5} \log n)$ edges satisfying $dist(x, y, H) \leqslant dist(x, y, G) + 2$.*

Next is a simple graph theoretic implication of the $+2$ distance preservers.

Corollary 5 *There are no graphs with girth (length of minimum cycle) five or more, which has $\Omega(n^{1.6})$ edges.*

Proof: Suppose G is a cycle of girth at least 5. Then, any $+2$ additive distance preserver of G must contain all the edges as we cannot afford to skip an edge. This means G must contain $O(n^{1.5} \log n)$ edges. □

Remark 6 *It is not possible to sparsify graphs while still preserving distances up to an additive error $+1$. A simple hindrance structure is a complete bipartite graph $K_{n,n}$.*

4 Preserving Large Distances

We see in this section another application of sampling in constructing sparse distance preserver for an n vertex graph $G = (V, E)$ that preserves distances between vertex pairs that are separated by a large distance of $\Omega(n)$.

Theorem 7 *For any undirected unweighted graph $G = (V, E)$ with n vertices and m edge, we can construct a sparse subgraph $H = (V, E_H \subseteq E)$ with $O(n \log n)$ edges such that for all $x, y \in V$ separated by distance at least $n/10$, we have*

$$dist(x, y, H) = dist(x, y, H).$$

Proof: Let $\mathcal{P}$ be set of all vertex pairs separated by a distance of at least $n/10$.

1. Initialize $H = (V, E_H)$ as empty graph, and let R be a uniformly random set of vertices of size $k = 40 \ln n$.

2. For each $v \in R$, compute the BFS tree rooted at v in G, say T_v, and add the $n-1$ edges of T_v to H. Thus, the graph H contains $O(n \log n)$ edges.

3. For each $(x, y) \in \mathcal{P}$, let

   - $Q_{x,y}$ be a shortest path from x to y in G, and
   - $\mathcal{E}_{x,y}$ be event that R contains zero vertices from $Q_{x,y}$.

4. Note that if event $\mathcal{E}_{x,y}$ does not hold then the distance from x to y in G and H is identical. This is because in such a case an (x, y) shortest path will contain a vertex from R, say r, due to which we will have:

$$\begin{aligned} dist(x, y, G) \leqslant dist(x, y, H) &\leqslant dist(x, r, H) + dist(r, y, H) \\ &= dist(x, r, G) + dist(r, y, G) \\ &= dist(x, y, G). \end{aligned}$$

5. For an $(x, y) \in \mathcal{P}$, we have:

$$\text{Prob}(\mathcal{E}_{x,y}) \leqslant \prod_{i=1}^{k} \left(\frac{n - n/10}{n}\right) \leqslant \left(1 - \frac{1}{10}\right)^{40 \ln n} \leqslant e^{-4 \ln n} = \frac{1}{n^4}.$$

6. By union bound, with probability at least $1 - |\mathcal{P}|/n^4$, for every $(x, y) \in \mathcal{P}$, the path $Q_{x,y}$ contains a vertex from R.

This proves the claim. □