COMS E6998-9: Algorithms for Massive Data (Spring'23) Sep 11, 2023

Lecture 2: Streaming for Graphs: Distance, Triangle Counting

Instructor: *Alex Andoni* Scribes: *Manuel Paez*

1 Distance

Consider an unweighted, undirected graph G with n nodes and m edges.
Problem: We want to stream over G by building a "spanner" graph H (where $H \subseteq G$) such that it preserves distances up to α factor

Algorithm 1 Algorithm for constructing spanner H

```
Initalize H = ∅
for each (i, j) ∈ G in the stream do
    if dist_H(i, j) > α then
        add (i, j) to H
    end if
end for
```

Claim 1. $dist_G(x, y) \leq dist_H(x, y) \leq \alpha \cdot dist_G(x, y)$

Proof. Suppose $\forall x, y \in G$. By construction of H, we will have a path from x to y that traverses k nodes:

$$x = v_0 \to v_1 \to \cdots \to v_k = y$$

With this, we can compute the distances between two nodes knowing there exists an alternate path in H

$$dist_H(x, y) \leq dist(v_0, v_1) + dist(v_1, v_2) + \cdots + dist(v_{k-1}, v_k)$$

$$\leq \alpha + \alpha + \cdots + \alpha \leq \alpha \cdot k = \alpha \cdot dist_G(x, y)$$

□

Claim 2 (Bollobas)**.** $|H| \leq O(n^{1+\frac{1}{k}})$ *where* $\alpha = 2k - 1$

Proof. Any cycle in H has length $\geq (\alpha + 1) + 1 = \alpha + 2 = 2k + 1$. With this, let's build an H' by deleting all nodes of $deg \leq n^{1/k}$. Thus, we deleted $\leq n \cdot n^{1/k}$ edges. Our graph H' will have:
1. all nodes with $deg > n^{1/k}$
2. minimum length cycle $\geq 2k + 1$ (Note: the minimum length cycle of a graph is the girth)
With this, let's fix r to be a node in H' such that a tree is formed where r is the root of the tree. This tree graph will have the property that all nodes up to depth k must differ from each other because if the nodes were similar it would create a cycle length $2k$ which cannot be true. By the kth level, we will have nodes

$$\geq (n^{1/k})^k (n^{1/k} + 1) > n$$

In exploring the tree, we cannot reach the k-th level of the tree and must stop earlier. However, if we stop, the leaves must have $deg = 1$ since they did not expand. Since these nodes would have been removed due to their degree, this implies that H' must be empty and therefore $|H| \leq O(n^{1+\frac{1}{k}})$ □

Question: does there exist a smaller spanner H?

Conjecture 3 (Erdős 63)**.** *There exists a graph G that has $|G| \geq \Omega(n^{1+\frac{1}{k}})$, girth $\geq \alpha + 2 = 2k + 1$, and size $\geq n^{1+\Omega(1/k)}$*

Theorem 4. *There exists a graph G that has girth $\geq 2k + 1$, size $\geq n^{1+\Omega(1/k)}$*

Theorem 5. *Erdős conjecture $\implies$ any α-spanner H must have size $|H| \geq \Omega(n^{1+\frac{1}{k}})$*

Proof. Fix graph G with the properties from Erdős conjecture. Assume the following:
1. if $H \subseteq G$, take any $(i, j) \in G \setminus H$
2. If H need not be $\subseteq G$
Let $\mathcal{F} = \{\text{all graphs } G' \subseteq G\}$. Consider $G' \in F$ such that it has a spanner. Let $H = \{\text{spanners of } G' \in F\}$.

Observation 6. *$\forall\ G_1, G_2 \in \mathcal{F}$, if $G_1 \neq G_2$, then their spanners must be different.*

To prove the observation, consider $(i, j) \in G_1 \setminus G_2$. With this, we have the following:

$$dist_{G_1}(i, j) = 1 \tag{1}$$

$$dist_{G_2}(i, j) \geq 2k \tag{2}$$

Given a fix H, then

$$dist_{G_1}(x, y) \leq dist_H(i, j) \leq \alpha \cdot dist_{G_1}(x, y) = \alpha = 2k = 1$$

The proof for the observation implies that $|\mathcal{H}| \geq |\mathcal{F}| = 2^m$ where $m = \#$ of edges in G. We have

$$|\mathcal{H}| \geq |\mathcal{F}| = 2^{\Omega(n^{1+1/k})}$$

Suppose the largest $H \in \mathcal{H}$ has size M, then

$$\implies |\mathcal{H}| \leq M \cdot \binom{n^2}{M} \leq n^2 n^{2M}$$

$$\implies 2M logn \geq \Omega(n^{1+\frac{1}{k}})$$

$$\implies M \geq \Omega\Big(n^{1+\frac{1}{k}} \cdot \frac{1}{log(n)}\Big)$$

□

Corollary 7. *fix any map $S : \mathcal{F} \to \{0, 1\}^M$, we can deduce all distances of G up to a factor α using $S(G)$*

The corollary implies that $M \geq log|\mathcal{H}| = \Omega(n^{1+\frac{1}{k}})$ bits are required to describe data structure

Definition 8. *distance oracle: data structure for undirected, unweighted graph G such that $\forall (i,j) \in G$, it can output an α-approx $dist_G(i,j)$ quickly*

Theorem 9 (Thorup-Zwick '05)**.** *Assume $\alpha = 2k-1$, then we can build a data structure with*
1. Size: $O(k \cdot n^{1+\frac{1}{k}})$
2. Query time: $O(k)$
3. Preprocessing $\equiv O(size + mlogn)$

2 Triangle Counting

Given a graph G, let $T = \#$ of triangles
Goal: Estimate T in the streaming model. For this, you are given t, assume that $T \geq t$, and use an $\alpha = 1 \pm \epsilon$ approximation
For S of three nodes, define a vector X such that $X_S = \#$ of edges inside S

Algorithm 2 Algorithm for constructing spanner H

Fix $K =$
Pick K random sets $S_1, S_2, ..., S_k$ where each subset S_i has three nodes
In the stream, compute all $X_{S_1}, ...X_{S_t}$
Compute $\hat{T} = \frac{\#\{i, X_{S_i}=3\}}{K} \cdot M$ where $M =$ total # of sets of size three ($\binom{n}{3}$)

Claim 10. $E[\hat{T}] = T$

Claim 11. $Pr[\hat{T} = T$ *up to* $1 \pm \epsilon] \geq 90\%$ *as long as K is large enough*