6.896 Sublinear Time Algorithms	March 13, 2007

Lecture 11

Lecturer: Ronitt Rubinfeld — *Scribe: Christina Sauper*

1 Estimating the Number of Connected Components

Given a graph $G(V, E)$ with max degree d and adjacency list representation and some ϵ, we want to give an additive estimate of the number of connected components to within ϵn.

1.1 Main Idea

Define:

$$n_u \equiv \text{number of nodes in } u\text{'s component, where } u \in V$$

Fact 1 *For any connected component $A \subseteq V$:*

$$\sum_{u \in A} \frac{1}{n_u} = \sum_{u \in A} \frac{1}{|A|} = 1$$

In addition, there are $\sum_{u \in V} \frac{1}{n_u}$ connected components.

Determining this value exactly takes $O(n^2)$ time, but we will estimate the sum and the values of n_u.

Define:

$$\hat{n}_u \equiv \min\left\{\text{nodes in } u\text{'s component}, \frac{2}{\epsilon}\right\}$$

$$\hat{c} = \sum_{u \in V} \frac{1}{\hat{n}_u}$$

Fact 2 *The error in estimating $\frac{1}{\hat{n}_u}$ is small.*

$$\left|\frac{1}{\hat{n}_u} - \frac{1}{n_u}\right| \leq \frac{\epsilon}{2}$$

Either $\hat{n}_u = n_u$ or $n_u > \hat{n}_u = \frac{2}{\epsilon}$. In the latter case, $\frac{\epsilon}{2} = \frac{1}{\hat{n}_u} \geq \frac{1}{n_u} \geq 0$. Therefore, the error is small, at most $\frac{\epsilon}{2}$.

Corollary 3 *$\frac{1}{\hat{n}_u}$ is a good estimate of connected components.*

$$\sum_{u \in V} \left|\frac{1}{n_u} - \frac{1}{\hat{n}_u}\right| \leq \frac{\epsilon n}{2}$$

$$c - \frac{\epsilon n}{2} \leq \frac{1}{\hat{n}_u} \leq c + \frac{\epsilon n}{2}$$

Fact 4 *We can compute $\hat{n}_u$ in $O(\frac{d}{\epsilon})$ time.*

Take $\frac{2}{\epsilon}$ steps of a BFS. If we see the entire connected component, set $\hat{n}_u = n_u = \frac{1}{\text{size}}$. Otherwise, $\hat{n}_u = \frac{2}{\epsilon}$.

Summing these $\hat{n}_u$ values yields a linear time algorithm. Now, we want to estimate this sum by estimating the average cluster size ($\sum_{u \in V} \frac{1}{\hat{n}_u}$) and multiplying by $|V|$.

1.2 Algorithm

Approx_Num_CC(G, ϵ)

Choose $r = O(\frac{1}{\epsilon^3})$ nodes $u_1 \ldots u_r$

$\forall u_i$ compute $\hat{n}_{u_i}$

Output $\tilde{c} = \frac{n}{r}\sum_{i=1}^{r}\frac{1}{\hat{n}_{u_i}}$

Runtime of this algorithm is $O(\frac{1}{\epsilon^3}\cdot\frac{d}{\epsilon}) = O(\frac{d}{\epsilon^4})$.

Theorem 5 $\Pr\left[|\tilde{c} - \hat{c}| \leq \frac{\epsilon}{2}n\right] \geq \frac{3}{4}$

Corollary 6 *Since* $|c - \tilde{c}| \leq |c - \hat{c}| + |\hat{c} - \tilde{c}|$ *and* $|c - \hat{c}| \leq \frac{\epsilon n}{2}$*:*

$$\Pr\left[|c - \tilde{c}| \leq \epsilon n\right] \geq \frac{3}{4}$$

Proof of theorem: We know upper and lower bounds on our estimated average cluster size:

$$\forall i \frac{\epsilon}{2} \leq \frac{1}{\hat{n}_i} \leq 1$$

Using Chernoff bounds, we can compute the error probability for the estimated cluster size:

$$\Pr\left[\left|\frac{1}{r}\sum_{1\leq i\leq r}\frac{1}{\hat{n}_{u_i}} - \text{Exp}\left[\frac{1}{\hat{n}_{u_i}}\right]\right| > \frac{\epsilon}{2}\text{Exp}\left[\frac{1}{\hat{n}_{u_i}}\right]\right] \leq \exp\left(-O(r\text{Exp}\left[\frac{1}{\hat{n}_{u_i}}\right]\cdot\left(\frac{\epsilon}{2}\right)^2)\right) \leq \frac{1}{4}$$

Here, using $r = \frac{c}{\epsilon^3}$ samples is good enough for constant c. The cutoff bound gets a better running time by bounding the maximum vs. minimum cluster sizes.

Likewise, we can see the error probability for the estimated sum:

$$\begin{array}{lccc} \Pr[& \left|\frac{n}{r}\cdot\sum\frac{1}{\hat{n}_{u_i}} - n\cdot\text{Exp}\left[\frac{1}{\hat{n}_{u_i}}\right]\right| & \leq & \epsilon\cdot\text{Exp}\left[\frac{n}{\hat{n}_{u_i}}\right]] \geq \frac{3}{4} \\ \Pr[& \left|\tilde{c} - \hat{c}(=\sum\frac{1}{\hat{n}})\right| & \leq & \epsilon\cdot\hat{c}(\leq n)] \geq \frac{3}{4} \end{array}$$

■

2 Minimum Spanning Tree

2.1 Definitions

Given a graph $G = (V, E)$ of degree $\leq d$, in adjacency list format and with edge weights $w_{ij} \in 1\ldots w \cup \infty$. We will assume the graph is connected; i.e., there is a minimum spanning tree of finite weight.

For a tree $T \subseteq E$:

$$\begin{aligned} w(T) &= \sum_{(ij)\in T} w_{ij} \\ M &= \min_{T \text{ spans } G} w(T) \end{aligned}$$

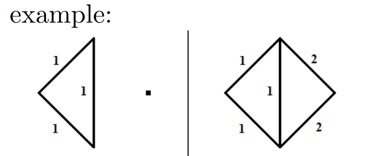
We will assume that all weights are positive and finite, therefore $n - 1 \leq M \leq \infty$.

2.2 Main Idea

Our goal is to output $\hat{M}$ such that $(1-\epsilon)M \le \hat{M} \le (1+\epsilon)M$. This is close to an ϵ-multiplicative estimate because $\frac{1}{1+\epsilon} \approx 1-\epsilon$.
Given a graph G:

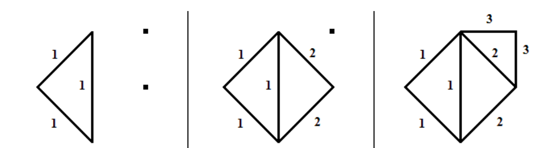
$$\begin{aligned} G^{(i)} &= \text{edges of } G \text{ which have weight at least } i \\ c^{(i)} &= \text{number of connected components in } G^{(i)} \end{aligned}$$

So the number of edges of weight at least k is $c^{(k-1)} - 1$.
For example:



$G^{(1)}$, $c^{(1)} = 2$	$G^{(2)}$, $c^{(2)} = 1$

$MST(G) = (n-1) + (c^{(1)} - 1) = n - 2 + c^{(1)} = 4$



$G^{(1)}$, $c^{(1)} = 3$	$G^{(2)}$, $c^{(2)} = 2$	$G^{(3)}$, $c^{(3)} = 1$

$MST(G) = (n-1) + (c^{(1)} - 1) + (c^{(2)} - 1) = n - 3 + c^{(1)} + c^{(2)} = 7$

Claim 7 $MST(G) = n - w + \sum_{1 \le i \le w-1} C^{(i)}$

Proof
Let α_i = the number of weight i edges in the MST.

Fact 8 *For any MST of G, α_i's are the same. Note that $\sum_{i=l+1}^{w} \alpha_i = c^{(l)} - 1$, and in particular $\sum_{i=1}^{w} \alpha_i = n-1$; $\alpha_w = c^{(w-1)} - 1$.*

$$\begin{aligned} MST(G) &= \sum_{i=1}^{w} i\alpha_i \\ &= \sum_{i=1}^{w} \alpha_i + \sum_{i=2}^{w} \alpha_i + \ldots + \alpha_w \\ &= n - 1 + c^{(1)} - 1 + c^{(2)} - 1 + \ldots + c^{(w-1)} - 1 \\ &= n - w + \sum_{i=1}^{w-1} c^{(i)} \end{aligned}$$

■

2.3 Algorithm

MST_Approx_Alg(G, ϵ, w)

for $i = 1 \ldots w-1$

$\hat{c}^{(i)}$ =Approx_Num_CC($G^{(i)}$, $\frac{\epsilon}{w}$)

Output $\hat{M} = n - w + \sum_{i=1}^{w-1} c^{(i)}$

Run time:

There are w calls to Approx_Num_CC (run time $O(d/(\frac{\epsilon}{w})^4)$), for an overall run time of $O(\frac{dw^5}{\epsilon^4})$. Because this running time depends on w, it is best when there is a good max to min ratio of edge weights.

Sketch of Proof $\forall i |\hat{c}^{(i)} - c^{(i)}| \le \frac{\epsilon}{w} n$ (with high enough probability) then $|M - \hat{M}| \le \epsilon n$.

Since $M > n$:

$(1-\epsilon)M \le \hat{M} \le M + \epsilon n \le M + \epsilon M = (1+\epsilon)M$

The lower bound is proved similarly. ■