CS246: Mining Massive Data Sets Problem Set 3

Name: Dat Nguyen
Date: 06/27/2019

By turning in this assignment, I agree by the Stanford honor code and declare that all of this is my own work.

1 Dead ends in PageRank computations (25 points)

(a) We have

$$\begin{aligned} w(\mathbf{r}') &= \sum_{i=1}^{n} r_i' \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} r_j \\ &= \sum_{j=1}^{n}\sum_{i=1}^{n} M_{ij} r_j \\ &= \sum_{j=1}^{n} r_j \sum_{i=1}^{n} M_{ij} \\ &= \sum_{j=1}^{n} r_j \\ &= w(\mathbf{r}) \end{aligned}$$

(b) We have

$$\begin{aligned} w(\mathbf{r}') &= \sum_{i=1}^{n} \left[\beta \sum_{j=1}^{n} M_{ij} r_j + \frac{(1-\beta)}{n}\right] \\ &= \beta \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} r_j + \sum_{i=1}^{n} \frac{1-\beta}{n} \\ &= \beta w(\mathbf{r}) + 1 - \beta \end{aligned}$$

We want $w(\mathbf{r}') = w(\mathbf{r})$, so

$$w(\mathbf{r}) = \beta w(\mathbf{r}) + 1 - \beta$$

This is satisfied if and only if $w(\mathbf{r}) = 1$ or $\beta = 1$ but we assume that $0 < \beta < 1$, therefore we conclude that $w(\mathbf{r}) = 1$.

(c)

2 Implementing PageRank and HITS (30 points)

(a) **PageRank Implementation [15 points]**

- Top 5 node ids with highest PageRank scores: 263, 537, 965, 243, 285.
- Bottom top 5 node ids with the lowest PageRank scores: 558, 93, 62, 424, 408.

(b) **HITS Implementation [15 points]**

- 5 node ids with highest hubbiness score: 840, 155, 234, 389, 472.
- 5 node ids with lowest hubbiness score: 23, 835, 141, 539, 889.
- 5 node ids with highest authority score: 893, 16, 799, 146, 473.
- 5 node ids with lowest authority score: 19, 135, 462, 24, 910.

3 Clique-Based Communities (25 points)

(a) Because there is a common factor i between every pair a and b so there is an edge between them. Therefore the set C_i for i be integer greater than 1 is a clique.

(b) Claim: C_i is a maximal clique if and only if i is prime number.
Proof:
Suppose C_i is a maximal clique. Suppose i is not a prime number, if we add a factor of i to the current clique, since every node in the current clique is divisible by that factor, the new set is also a clique. This contradicts the assumption that C_i is a maximal clique.
Now suppose that i is prime number. If we add a new node j to the current clique to make a bigger clique, then j must be divisible by i. Therefore j must be in current clique which shows that there is no way to expand the clique. (q.e.d)

(c) Assume that the unique largest clique is C'. Let $i_1, i_2, \ldots$ be the nodes of C' from the smallest node to the largest node. Because the smallest node in G is 2 we have $i_1 \geq 2$. Since i_1 and i_2 have a common factor other than 1, $i_2 \geq i_1 + 2$. Following that argument we also have $i_{j+1} \geq i_j + 2$ for any j and $j + 1$ be the indexes of nodes in C'. Therefore the series of nodes $i_1, i_2, \ldots$ can only have as many elements as C_2, only when C' is C_2, which concludes that C_2 is the unique largest clique.

4 Dense Communities in Networks (20 points)

(a) i Let $B(S) = S \setminus A(S)$, suppose the contrary

$$\begin{aligned} & |A(S)| < \frac{\epsilon}{1+\epsilon}|S| \\ \Leftrightarrow \quad & |S| - |A(S)| \geq |S| - \frac{\epsilon}{1+\epsilon}|S| \\ \Leftrightarrow \quad & |B(S)| \geq \frac{1}{1+\epsilon}|S| \end{aligned}$$

And we also have $B(S) = \{j \in S | \deg_S(j) > 2(1+\epsilon)\rho(S)\}$. Therefore the sum of degree of all nodes in $B(S)$ is bounded by

$$\begin{aligned} \sum_{j \in B(S)} \deg_S(j) &> |B(S)| 2(1+\epsilon)\rho(S) \\ &\geq \frac{1}{1+\epsilon}|S| 2(1+\epsilon)\frac{|E[S]|}{|S|} \\ &= 2|E[S]| \\ &= \sum_{i \in S} \deg_S(i) \end{aligned}$$

Which is a contradiction since $B(S)$ is a subset of S. So we conclude that $|A(S)| \geq \frac{\epsilon}{1+\epsilon}|S|$

ii After one iteration the number of elements remaining in S is $B(S) < \frac{1}{1+\epsilon}|S|$. Because $|S|$ always shrinks after every iteration, after finite number of iterations S will become the empty set. Let the number of iterations required be m, since after m - 1 iterations $|S| \geq 1$ (because $|S|$ decreases by at least 1 after every iteration), we have

$$\begin{aligned} 1 &\leq \left(\frac{1}{1+\epsilon}\right)^{m-1} n \\ (1+\epsilon)^{m-1} &\leq n \\ m - 1 &\leq \log_{1+\epsilon} n \\ m &\leq \log_{1+\epsilon} n + 1 \end{aligned}$$

Therefore m = O($\log_{1+\epsilon} n$)

(b) i Suppose that there exists a node $v' \in S^*$ that $\deg_{S^*}(v') < \rho(G)$, if we remove v'

from S^* we get the new set S' and the density

$$\begin{aligned}\frac{|E[S']|}{|S'|} &= \frac{|E[S^*]| - \deg_{S^*}(v')}{|S^*| - 1} \\ &> \frac{|E[S^*]| - \rho^*(G)}{|S^*| - 1} \\ &= \frac{|E[S^*]| - \frac{|E[S^*]|}{|S^*|}}{|S^*| - 1} \\ &= \frac{|E[S^*]|(|S^*| - 1)}{|S^*|(|S^*| - 1)} \\ &= \frac{|E[S^*]|}{S^*}\end{aligned}$$

Which contradicts with the fact that S^* is the densest subgraph of G.

ii Because in the first iteration $S = V$, we have $\deg_S(v) \geq \deg_{S^*}(v)$. This together with the fact that $\deg_{S^*}(v) \geq \rho^*(G)$ implies that $\deg_S(v) \geq \rho^*(G)$. Because $v \in A(S)$, we have $\deg_S(v) \leq 2(1+\epsilon)\rho(S)$. Combine 2 facts above, we arrive at $2(1+\epsilon)\rho(S) \geq \rho^*(G)$.

iii Since after every iteration $\rho(\tilde{S})$ never decreases, we have

$$\rho(\tilde{S}) \geq \rho(S) \geq \frac{1}{2(1+\epsilon)}\rho^*(G)$$