CSE547: Machine Learning for Big Data
Homework 3

January Shen
May 13, 2019

Answer to Question 1(a)

We want to prove that w(r') = w(r)

$$r' = M * r$$

$$w(r') = \sum_{i=1}^{n} r'_i = \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} * r_j$$

Because both sum are finite sum, we can interchange the summation, s.t.

$$w(r') = \sum_{j=1}^{n} \sum_{i=1}^{n} M_{ij} * r_j$$

We know that $\sum_{i=1}^{n} M_{ij} = 1$ for each j since there is no dead end. Then we get

$$w(r') = \sum_{j=1}^{n} 1 * r_j = \sum_{j=1}^{n} r_j = w(r)$$

Answer to Question 1(b)

$$r'_i = \beta \sum_{j=1}^{n} M_{ij} * r_j + (1 - \beta)/n$$

$$w(r') = \beta \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} * r_j + (1 - \beta)$$

From Question 1(a) we know that $\sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} * r_j = \sum_{j=1}^{n} *r_j$, so we get $w(r') = \beta \sum_{j=1}^{n} *r_j + (1 - \beta) = \beta * w(r) + (1 - \beta)$.

Let $w(r') = w(r) = x$. By solving the equation $x = \beta * x + 1 - \beta$, we get that the equation holds when $x = 1$.

As such, under the circumstances that $w(r) = 1$ and $w(r') = 1$, $w(r) = w(r')$.

Answer to Question 1(c)

We can write r'_i as follows, where D is the set of dead nodes:

$$r'_i = \beta(\sum_{j \notin D}^{n} M_{ij} * r_j + \sum_{j \in D}^{n} 1/n * r_j) + (1-\beta)/n$$

By applying summation over i we get $w(r')$:

$$w(r') = \beta(\sum_{i=1}^{n} \sum_{j \notin D}^{n} M_{ij} * r_j + \sum_{i=1}^{n} \sum_{j \in D}^{n} r_j/n) + (1-\beta)$$

From Question 1(a) we know that $\sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} * r_j = \sum_{j=1}^{n} * r_j$, so

$$w(r') = \beta(\sum_{j \notin D}^{n} r_j + \sum_{j \in D}^{n} r_j) + (1-\beta) = \beta * w(r) + (1-\beta)$$

Assume $w(r) = 1$, we get that $w(r') = 1$.

Answer to Question 2(a)

graph-full.txt
The top 1 Node is 263 with PageRank score of 0.002020
The top 2 Node is 537 with PageRank score of 0.001943
The top 3 Node is 965 with PageRank score of 0.001925
The top 4 Node is 243 with PageRank score of 0.001853
The top 5 Node is 285 with PageRank score of 0.001827

The bottom 1 Node is 558 with PageRank score of 0.00032860
The bottom 2 Node is 93 with PageRank score of 0.00035136
The bottom 3 Node is 62 with PageRank score of 0.00035315
The bottom 4 Node is 424 with PageRank score of 0.00035482
The bottom 5 Node is 408 with PageRank score of 0.00038780

Answer to Question 2(b)

graph-full.txt
The top 1 hubbiness node is 840 with hubbiness score 1.000000
The top 2 hubbiness node is 155 with hubbiness score 0.949962
The top 3 hubbiness node is 234 with hubbiness score 0.898665
The top 4 hubbiness node is 389 with hubbiness score 0.863417
The top 5 hubbiness node is 472 with hubbiness score 0.863284
The bottom 1 hubbiness node is 23 with hubbiness score 0.042067
The bottom 2 hubbiness node is 835 with hubbiness score 0.057791
The bottom 3 hubbiness node is 141 with hubbiness score 0.064531
The bottom 4 hubbiness node is 539 with hubbiness score 0.066027
The bottom 5 hubbiness node is 889 with hubbiness score 0.076784

The top 1 authority node is 893 with authority score 1.000000
The top 2 authority node is 16 with authority score 0.963557
The top 3 authority node is 799 with authority score 0.951016
The top 4 authority node is 146 with authority score 0.924670
The top 5 authority node is 473 with authority score 0.899866
The bottom 1 authority node is 19 with authority score 0.056083
The bottom 2 authority node is 135 with authority score 0.066539
The bottom 3 authority node is 462 with authority score 0.075442
The bottom 4 authority node is 24 with authority score 0.081712
The bottom 5 authority node is 910 with authority score 0.085717

Answer to Question 3(a)

i
$|S|$:= the number of node in subset S.
$|E[S]|$:= the number of edge in subset S.

Prove that $A(S) \geq \frac{\epsilon}{1+\epsilon}|S|$.

We know that $2|E[S]| = \sum_{v \in S} deg_S(v) \geq \sum_{v \in S \setminus A(S)} deg_S(v)$

We also know that $\sum_{v \in A(S)} deg_{A(S)}(v) \leq |A(S)| * 2(1+\epsilon)\rho(S)$.

By subtracting both side from $\sum_{v \in S} deg_S(v) = |S| * 2\rho(S)$, we get $\sum_{v \in S \setminus A(S)} deg_S(v) > |S \setminus A(S)| * 2(1+\epsilon) * \rho(S)$

Finally, we get

$$2|E[S]| > |S \setminus A(S)| * 2(1+\epsilon)\rho(S)$$

Since all the elements are non-negative, we can write the equation below:

$$\frac{|E[S]|}{(1+\epsilon)\rho(S)} > |S \setminus A(S)|$$

$$\frac{|E[S]|}{(1+\epsilon)\rho(S)} = \frac{|S|\rho(S)}{(1+\epsilon)\rho(S)} > |S| - |A(S)|$$

$$A(S) > \frac{\epsilon}{1+\epsilon}|S|$$

ii
We know that $|S|$ shrinks to at least $\frac{1}{1+\epsilon}|S|$ in each iteration, so with n nodes, it takes at most $log_{1+\epsilon}(n)$ iterations to terminate the algorithm.

Answer to Question 3(b)

i. Prove that for any $v \in S^*$, $deg_{S^*}(v) \geq \rho^*(G)$.

Assume that S^* is the densest subgraph of G, and there exists an $v \in S^*$ where $deg_{S^*}(v) < \rho^*(G)$.

Then $\rho|S^* \setminus (v)| = \frac{|E[S^*]| - deg_{S^*}(v)}{|S^*|-1} \geq \frac{|E[S^*]| - \rho^*(G)}{|S^*|-1} = \frac{|E[S^*]| - \rho^*(S^*)}{|S^*|-1} = \frac{|E[S^*]|}{|S^*|} = \rho|S^*|$

This contradicts the assumption that S^* is the densest subgraph of G. We proved that for any $v \in S^*$, $deg_{S^*}(v) \geq \rho^*(G)$.

ii. Prove that $2(1+\epsilon)\rho(S) \geq \rho^*(G)$.

In the first interation, if there exists a node $v \in S^* \cap A(S)$, then we have: $2(1+\epsilon)\rho(S) \geq deg_{S^*}(v)$. From (i) we know that $deg_{S^*}(v) \geq \rho^*(G)$, so we proved that $2(1+\epsilon)\rho(S) \geq deg_{S^*}(v) \geq \rho^*(G)$.

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

We know that $\rho(\tilde{S}) \geq \rho(S)$ from the algorithm's definition. From ii, we also know that $\rho(S) \geq \frac{1}{2(1+\epsilon)}\rho^*(G)$. We thus can conclude that $\rho(\tilde{S}) \geq \frac{1}{2(1+\epsilon)}\rho^*(G)$

Answer to Question 4(a)

1. Prove that $L = \sum_{\{i,j\}\in E}(e_i - e_j)(e_i - e_j)^T$

From observation we see that
$A = \sum_{\{i,j\}\in E}(e_i + e_j)(e_i + e_j)^T - e_i e_i^T - e_j e_j^T$
$D = \sum_{\{i,j\}\in E} e_i e_i^T + e_j e_j^T$

$L = D - A = 2(\sum_{\{i,j\}\in E} e_i e_i^T + e_j e_j^T) - \sum_{\{i,j\}\in E}(e_i + e_j)(e_i + e_j)^T = \sum_{\{i,j\}\in E} e_i e_i^T + e_j e_j^T - e_i e_j^T - e_j e_i^T = \sum_{\{i,j\}\in E}(e_i - e_j)(e_i - e_j)^T$.

2. Prove that for any vector $x \in R^n$, it holds that $x^T Lx = \sum_{\{i,j\}\in E}(x_i - x_j)^2$.

$x^T Lx = x^T \sum_{\{i,j\}\in E}(e_i - e_j)(e_i - e_j)^T x = \sum_{\{i,j\}\in E}(x_i - x_j)(x_i - x_j) = \sum_{\{i,j\}\in E}(x_i - x_j)^2$.

3. $x_S^T L x_S = c$. Show NCUT(S) in c.

From 2 we know that $x_S^T L x_S = \sum_{i\in S, j\notin S}(\sqrt{\frac{vol(\bar{S})}{vol(S)}} + \sqrt{\frac{vol(S)}{vol(\bar{S})}})^2 + \sum_{i\notin S, j\in S}(\sqrt{\frac{vol(S)}{vol(\bar{S})}} + \sqrt{\frac{vol(\bar{S})}{vol(S)}})^2$.

Because $\bar{S} = V \setminus S$:

$c = \sum_{\{i,j\}\in E}(\sqrt{\frac{vol(S)}{vol(\bar{S})}} + \sqrt{\frac{vol(\bar{S})}{vol(S)}})^2 = \sum_{\{i,j\}\in E}(\frac{vol(S)^2 + vol(\bar{S})^2}{vol(S)vol(\bar{S})}) = \sum_{\{i,j\}\in E}(\frac{(2m)^2}{vol(S)vol(\bar{S})} - 2)$.

$NCUT(S) = \frac{cut(S)}{vol(S)} + \frac{cut(\bar{S})}{vol\bar{S}} = 2m\frac{cut(S)}{vol(S)vol(\bar{S})} = \frac{c+2}{2m}cut(S)$

4. Prove that $x_S^T De = 0$

$x_S^T De = \sum_{i\in S} d_i \sqrt{\frac{vol(\bar{S})}{vol(S)}} - \sum_{j\in\bar{S}} d_j \sqrt{\frac{vol(S)}{vol(\bar{S})}} = \sum_{i\in S} d_i \sqrt{\frac{\sum_{i\in\bar{S}} d_i}{\sum_{i\in S} d_i}} - \sum_{j\in\bar{S}} d_j \sqrt{\frac{\sum_{j\in S} d_j}{\sum_{j\in\bar{S}} d_j}} = 0$.

5. Prove that $x_S^T D x_S = 2m$

$x_S^T D x_S = \sum_{i\in S} d_i x_S^2 = \sum_{i\in S} d_i(\frac{vol(\bar{S})}{vol(S)}) + \sum_{i\in\bar{S}} d_i(\frac{vol(S)}{vol(\bar{S})}) = \sum_{i\in S} d_i(\frac{\sum_{i\in\bar{S}} d_i)}{\sum_{i\in S} d_i}) + \sum_{i\in\bar{S}} d_i(\frac{\sum_{i\in S} d_i}{\sum_{i\in\bar{S}} d_i}) = \sum_{i\in V} d_i = \sum_{i\in V}\sum_{j=1}^n A_{ij} = 2m$.

Answer to Question 4(b)

Answer to Question 4(c)

Prove that $Q(y) = \frac{1}{2m}(-2 * cut(S) + \frac{1}{m}vol(S) * vol(\bar{S}))$

$$Q(y) = \frac{1}{2m}\sum_{i,j=1}^{n}[A_{ij} - \frac{d_i d_j}{2m}]\delta(y_i, y_j) = \frac{1}{m}\sum_{i,j\in S}[A_{ij} - \frac{d_i d_j}{2m}] = \frac{1}{m}[(vol(S) - cut(S)) - \frac{\sum_{i\in S} d_i \sum_{j\in S} d_j}{2m}]$$

$$= \frac{1}{m}(vol(S) - cut(S) - \frac{vol(S)^2}{2m} = \frac{1}{m}[\frac{2m * vol(S)}{2m} - cut(S) - \frac{vol(S)^2}{2m}]$$

$$= \frac{1}{m}[\frac{vol(S)^2 + vol(\bar{S})vol(S)}{2m} - cut(S) - \frac{vol(S)^2}{2m}] = \frac{1}{m}[\frac{vol(\bar{S})vol(S)}{2m} - cut(S)]$$

$$= \frac{1}{2m}(-2 * cut(S) + \frac{1}{m}vol(\bar{S})vol(S))$$