Question 1(a), Homework 3, CS246

$w(r^{'}) = \sum_i r^{'}_i = \sum_i \sum_j M_{ij} r_j = \sum_j r_j \sum_i M_{ij} = \sum_j r_j 1 = w(r)$

Iff $w(r) = 1$.

$w(r') = \sum_i r_i' = \sum_i(\beta \sum_j M_{ij} r_j + (1-\beta)\frac{1}{n})$

$= 1 - \beta + \beta \sum_j r_j \sum_i M_{ij} = 1 - \beta + \beta \sum_j r_j = 1 - \beta + \beta w(r) = 1$

$r_i = \beta \sum_i M_{ij} r_j + \sum_{j \not\epsilon D} (1-\beta) r_j \frac{1}{n} + \sum_{j \epsilon D} r_j \frac{1}{n} = \beta \sum_j M_{ij} r_j + \frac{(1-\beta)}{n} + \frac{\beta}{n} \sum_{j \epsilon D} r_j$

$w(r^{'}) = \sum_i r_i^{'} = \sum_i (\beta \sum_j M_{ij} r_j + \sum_{j \not\epsilon D} (1-\beta) r_j \frac{1}{n} + \sum_{j \epsilon D} r_j \frac{1}{n}) = \beta \sum_j r_j \sum_i M_{ij} + \sum_i ...$

$= \beta \sum_{j \not\epsilon D} r_j + \sum_i (\sum_j r_j \frac{1}{n} - \beta \frac{1}{n} \sum j \not\epsilon D r_j) = 1 + \beta \sum_{j \not\epsilon D} r_j - \beta \sum_{j \not\epsilon D} r_j = 1$

- Top 1st node ID: 263
- Top 2nd node ID: 537
- Top 3rd node ID: 965
- Top 4th node ID: 243
- Top 5th node ID: 285

- Bottom 1st node ID: 558
- Bottom 2nd node ID: 93
- Bottom 3rd node ID: 62
- Bottom 4th node ID: 424
- Bottom 5th node ID: 408

- Top 1st Hubbines node ID: 840
- Top 2nd Hubbines node ID: 155
- Top 3rd Hubbines node ID: 234
- Top 4th Hubbines node ID: 389
- Top 5th Hubbines node ID: 472

- Bottom 1st Hubbines node ID: 23
- Bottom 2nd Hubbines node ID: 835
- Bottom 3rd Hubbines node ID: 141
- Bottom 4th Hubbines node ID: 539
- Bottom 5th Hubbines node ID: 889

- Top 1st Authority node ID: 893
- Top 2nd Authority node ID: 16
- Top 3rd Authority node ID: 799
- Top 4th Authority node ID: 146
- Top 5th Authority node ID: 473

- Bottom 1st Authority node ID: 19
- Bottom 2nd Authority node ID: 135
- Bottom 3rd Authority node ID: 462
- Bottom 4th Authority node ID: 24
- Bottom 5th Authority node ID: 910

If there are 0 or 1 element in C_i, it is clique by a definition. If there are at least two elements, any pair of elements that share a common factor i, so any pair is connected, so C_i is a clique.

Condition: $i <= 10^6$ and i must be a prime $\Leftrightarrow C_i$ is a maximum clique.

If $i >= 10^6$, then C_i is empty clique and if we add any node to C_i will produce a 1-clique.

If $i <= 10^6$ but is not prime. We can denote j as a factor of i, where $1 < j < i$. Node j is not yet in C_i, it have an edge to every member of C_i, because j is a factor of i.

Final: If $i <= 10^6$ and prime there is no node outside of C_i that have edge to i. If we want to check that we can denote node called j then i and j should have any other common factor as 1, but that could not be the case since i is prime. If j have i as factor, than j is a multiple of i so that means it already is in C_i.

Lemma: Largest possible clique have 500000 elements, because a pair of consecutive integers are always coprime. We can only choose one of each $\{2, 3\}, \{4, 5\}, ..., \{999998, 999999\}, \{1000000\}$ to form a clique. Therefore we can have most 500000 elements.

Uniqueness: By lemma the largest clique include either 2 or 3, if we pick 3, the best we can do is to include all multiplies of 3, which will form us C_3, which have less than 500000 elements.

C_2 is unique largest clique.

Note to understand the problem: $\rho(S) = \frac{E|[S]|}{|S|}$ is half of the average degree.

Note to understanding pseudocode: $\widetilde{S}$ is to keep the best S.

$i.$

$2E\,|[S]| = \sum_{i \epsilon S} deg_s(i)$ (Each edge in $E[S]$ contributes to 2 among all $deg_s(i)$.)

$>= \sum_{i \epsilon S \setminus A} deg_s(i)$ ($S \setminus A \subset S$)

$> \sum_{i \epsilon S \setminus A} 2(1+\epsilon)\frac{E|[S]|}{|S|}$ (from the definition of $A(S)$ in the pseudocode)

$= |S \setminus A|\, 2(1+\epsilon)\frac{E|[S]|}{|S|}$

Therefore,

$\frac{1}{1+\epsilon}\,|S| > |S \setminus A| = |S| - |A|$ ($A \subset S$)

$\Rightarrow |A| > \frac{\epsilon}{1+\epsilon}\,|S|$

$ii.$

Since $\frac{1}{1+\epsilon}\,|S| > |S \setminus A|$ after every iteration, $|S|$ decreases by at least $1+\epsilon$ times, so there are $O(log_{1+\epsilon}(n))$ iterations.

$i.$

Let say $deg_{s*}(v) < p^*(G)$. Then $\frac{deg_{S*}(v)}{1} < \frac{|E[S^*]|}{|S^*|}$.

Then $\rho(S^* \setminus v) = \frac{|E[S^*]| - deg_{S*(v)}}{|S^*|-1} > \rho(S^*)$, so contradiction.

$ii.$

$2(1+\epsilon)\rho(S) >= deg_S(v)$ because $(v \in A)$

$>= \rho^*(G)$ (by $i.$, since $v \in S^*$)

$iii.$

$\rho^*(\widetilde{S}) >= \rho(S)$ (because of the algorithm)

$>= \frac{1}{2(1+\epsilon)}\rho^*(G)$ (by $ii.$)

Information sheet
CS246: Mining Massive Data Sets

Assignment Submission Fill in and include this information sheet with each of your assignments. This page should be the last page of your submission. Assignments are due at 11:59pm and are always due on a Thursday. All students (SCPD and non-SCPD) must submit their homework via Gradescope (http://www.gradescope.com). Students can typeset or scan their homework. Make sure that you answer each (sub-)question on a separate page. That is, one answer per page regardless of the answer length. Students also need to upload their code on Gradescope. Put all the code for a single question into a single file and upload it.

Late Homework Policy Each student will have a total of *two* late periods. *Homework are due on Thursdays at 11:59pm PT and one late period expires on the following Monday at 11:59pm PT*. Only one late period may be used for an assignment. Any homework received after 11:59pm PT on the Monday following the homework due date will receive no credit. Once these late periods are exhausted, any assignments turned in late will receive no credit.

Honor Code We strongly encourage students to form study groups. Students may discuss and work on homework problems in groups. However, each student must write down their solutions independently, i.e., each student must understand the solution well enough in order to reconstruct it by him/herself. Students should clearly mention the names of all the other students who were part of their discussion group. Using code or solutions obtained from the web (GitHub/Google/previous year's solutions etc.) is considered an honor code violation. We check all the submissions for plagiarism. We take the honor code very seriously and expect students to do the same.

Your name:

Email: **SUID:**

Discussion Group:

I acknowledge and accept the Honor Code.

(Signed)