CS 660 - Data Science at Scale

Assignment 1: PageRank

Group: Eric Jiang, Joonsoo Park, Franco Pettigrosso, McWelling Todman

1 Read paper by Bryan and Leiseand answer the exercises in the paper

- Exercise 1. Suppose the people who own page 3 in the web of Figure 1 are infuriated by the fact that its importance score, computed using formula (2.1), is lower than the score of page 1. In an attempt to boost page 3s score, they create a page 5 that links to page 3; page 3 also links to page 5. Does this boost page 3s score above that of page 1?
- Exercise 3. Add a link from page 5 to page 1 in the web of Figure 2. The resulting web, considered as an undirected graph, is connected. What is the dimension of V1(A)?
- Exercise 4. In the web of Figure 2.1, remove the link from page 3 to page 1. In the resulting web page 3 is now a dangling node. Set up the corresponding substochastic matrix and find its largest positive (Perron) eigenvalue. Find a non-negative Perron eigenvector for this eigenvalue, and scale the vector so that components sum to one. Does the resulting ranking seem reasonable?
- Exercise 5. Prove that in any web the importance score of a page with no backlinks is zero.

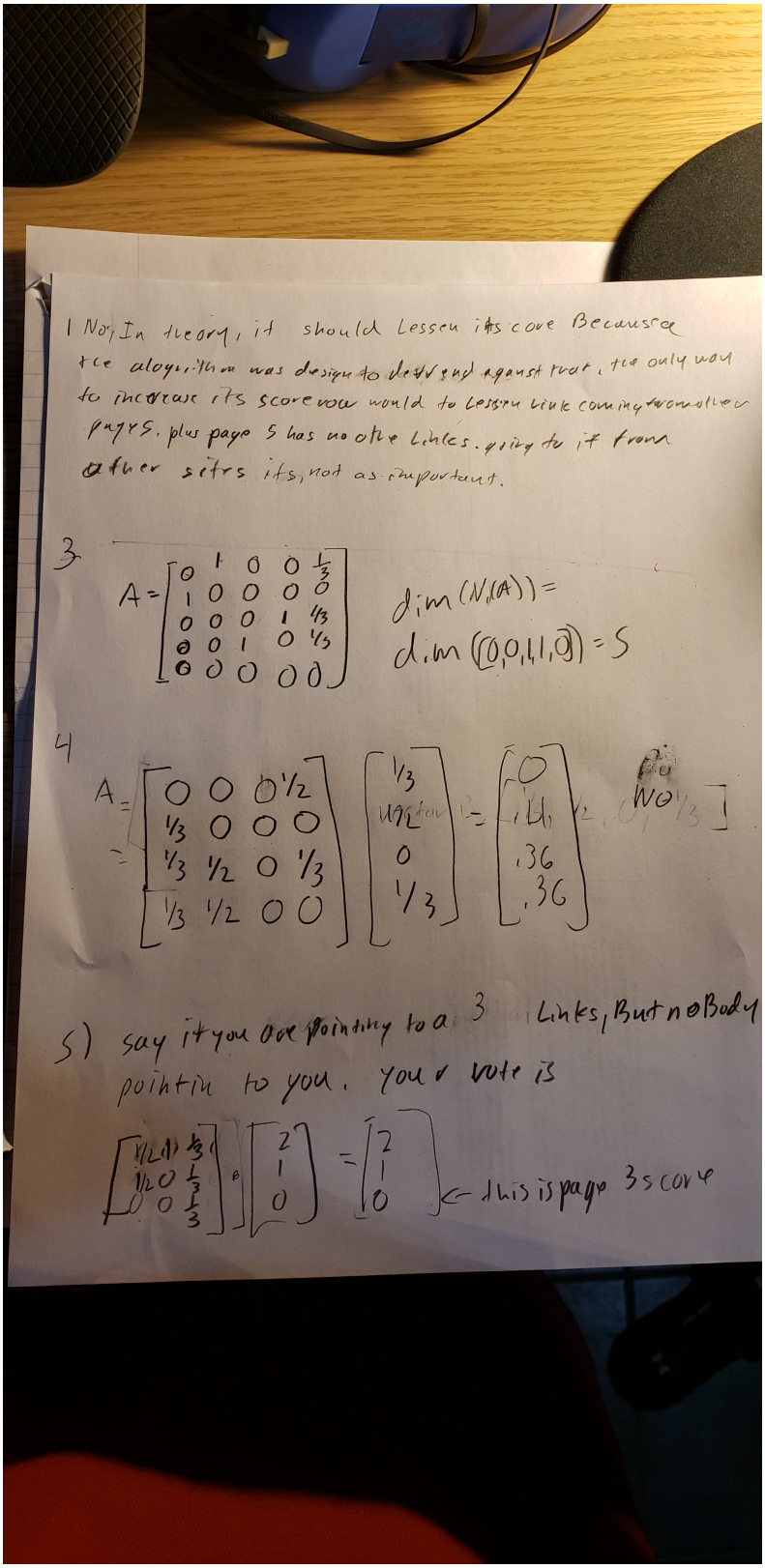


Figure 1: Answers to Exercises 1,3,4,5

- Exercise 2. Construct a web consisting of three or more subwebs and verify that dim(V1(A)) equals (or exceeds) the number of the components in the web.

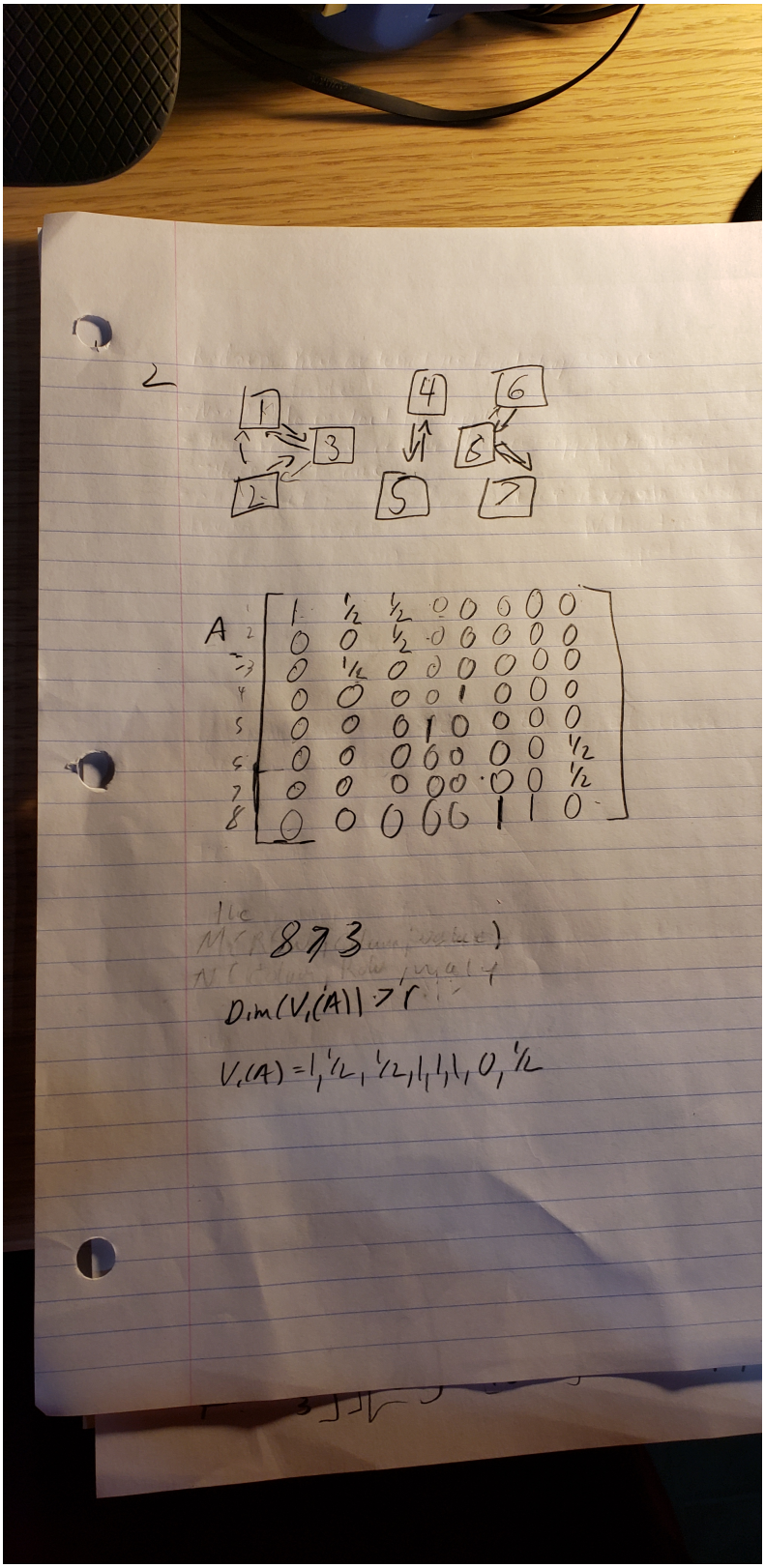


Figure 2: Answers to Exercise 2

- Exercise 6. Implicit in our analysis up to this point is the assertion that the manner in which the pages of a web W are indexed has no effect on the importance score assigned to any given page. Prove this, as follows: Let W contains n pages, each page assigned an index 1 through n, and let A be the resulting link matrix. Suppose we then transpose the indices of pages i and j (so page i is now page j and vice-versa). Let A be the link matrix for the relabelled web.

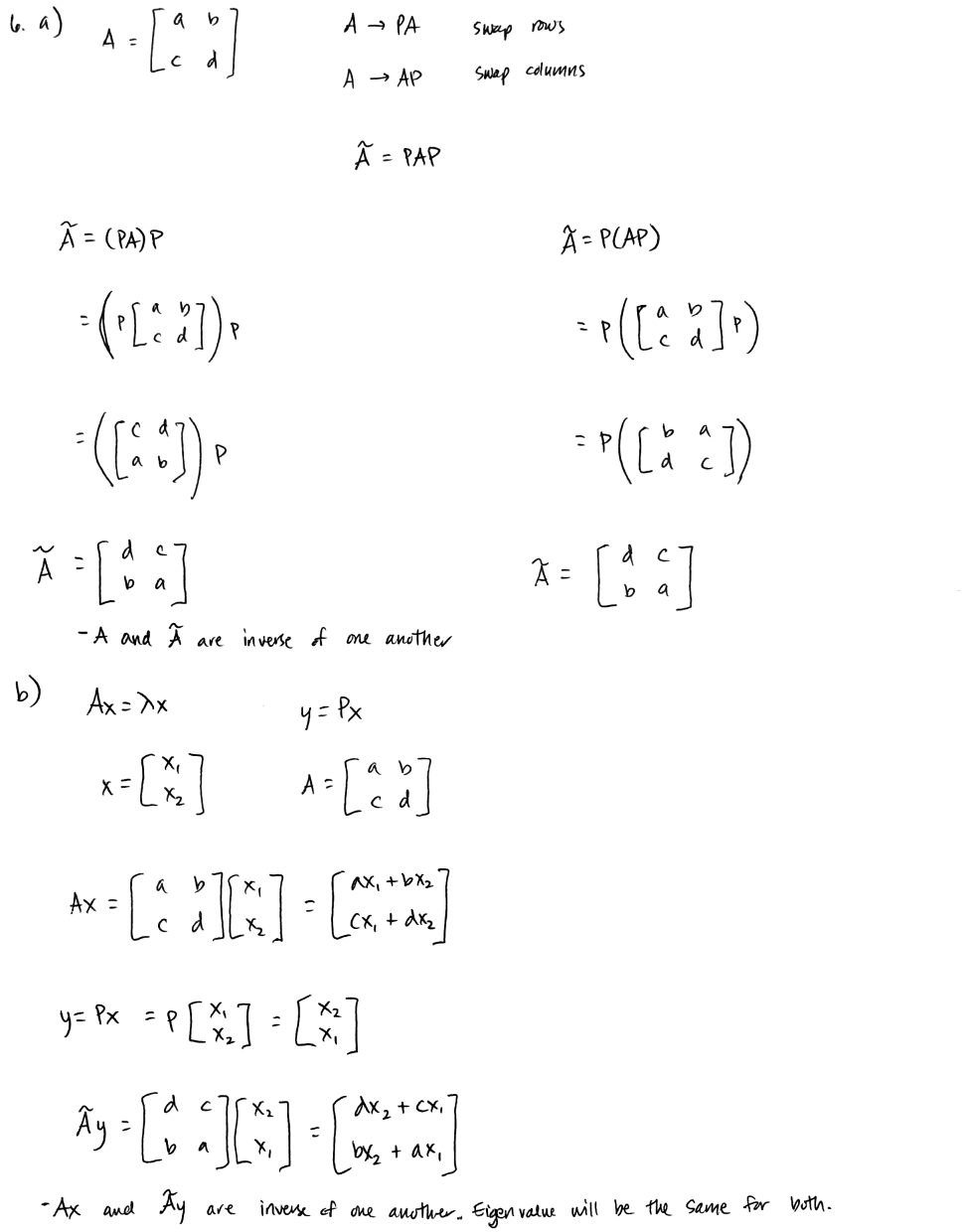


Figure 3: Answers to Exercise 6 parts 1 and 2

  - Part 3: This shows that transposing the indices of any two pages leaves the importance scores unchanged because the eigenvector is the same for both the original matrix and the new matrix. So, whether it is at it's original position or at the inverted position, the eigenvector will lead to the same importance score. This will apply to any permutation of the page indices. As long as the changes are just transposing or transversing of the

indices, then the results should remain consistent.

Exercise 7. Prove that if A is an n × n column-stochastic matrix and $0 \leq m \leq 1$, then $M = (1-m)A + mS$ is also a column-stochastic matrix.

- given that A is column stochastic, each column sums to 1. Therefore by multiplying A by $1-m$, each column will now sum to $1-m$ rather than 1.
- the expression mS corresponds to a $n \times n$ matrix in which each element is $\frac{m}{n}$, where n is the span of M. Each element is in essence, a fraction that if added to itself n times, would result in the value m, which is conveniently what we eliminated from each column of A in the previous bullet.
- finally, we can be confident $M = (1-m)A + mS$ is a column stochastic matrix because by adding the matrix mS, we are adding $\frac{m}{n}$ to each of the n elements in each column of A, thus undoing the reduction that took place in the first bullet.

- Exercise 8. Show that the product of two column-stochastic matrices is also column-stochastic.

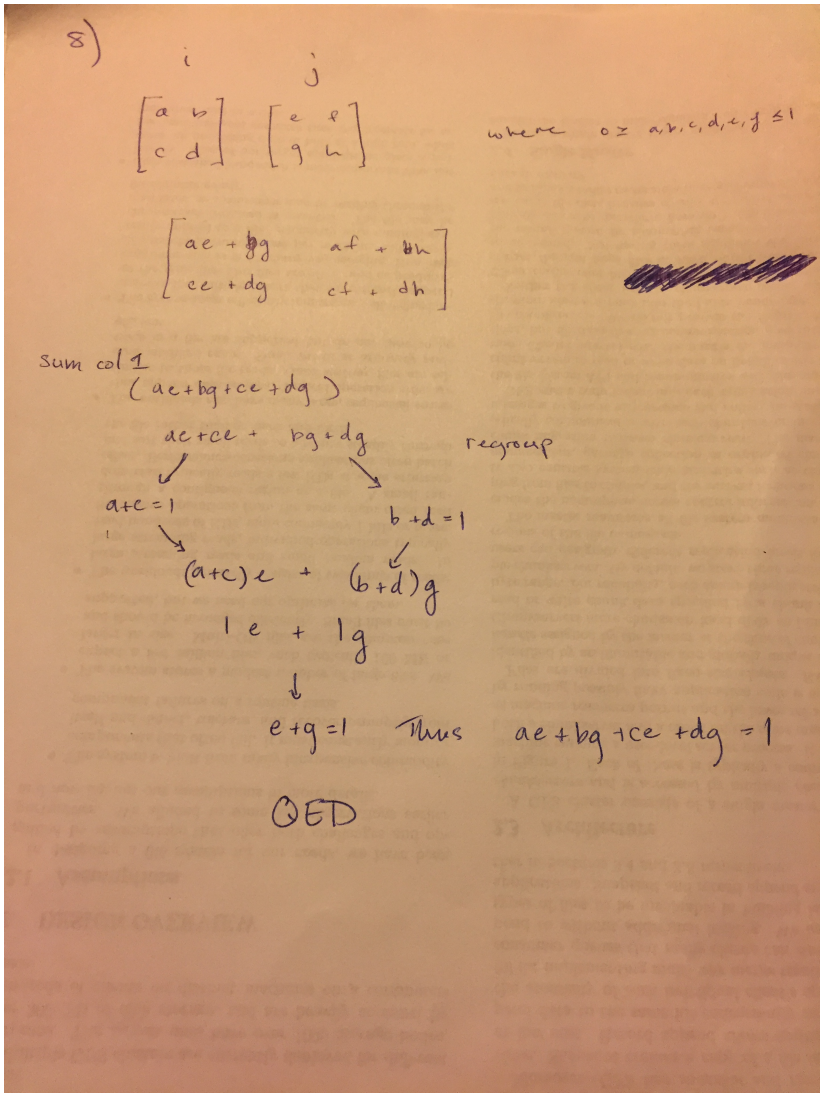


Figure 4: Answers to Exercise 8

- Exercise 9. Show that a page with no backlinks is given importance score $\frac{m}{n}$ by formula (3.2).
  - see exercise9.py file
- Exercise 10. Suppose that A is the link matrix for a strongly connected web of n pages (any page can be reached from any other page by following a finite number of links). Show that $dim(V1(A)) = 1$ as follows. Let $(A^k)_{ij}$ denote the (i, j)-entry of A^k.
  - Note that page i can be reached from page j in one step if and only $A_{ij} > 0$ (since $A_{ij} > 0$ means theres a link from j to i!) Show that $(A^2)_{ij} > 0$ if and only if page i can be reached from page j in exactly two steps. Hint: $(A^2)_{ij} = \Sigma_k A_{ik} A_{kj}$; all A_{ij} are non-negative, so $(A^2)ij > 0$ implies that for some k both A_{ik} and A_{kj} are positive.
  - Show more generally that $(A^p)_{ij} > 0$ if and only if page i can be reached from page j in EXACTLY p steps.
  - Argue that $(I + A + A^2 + ... + A^p)_{ij} > 0$ if and only if page i can be reached from page j in p or fewer steps (note $p = 0$ is a legitimate choiceany page can be reached from itself in zero steps!)
  - Explain why $I + A + A^2 + + A^{n-1}$ is a positive matrix if the web is strongly connected.
  - Use the last part (and Exercise 8) so show that $B = \frac{1}{n}(I + A + A^2 + + A^{n-1})$ is positive and column-stochastic (and hence by Lemma 3.2, $dim(V1(B)) = 1$).
  - Show that if $x \in V1(A)$ then $x \in V1(B)$. Why does this imply that $dim(V1(A)) = 1$?

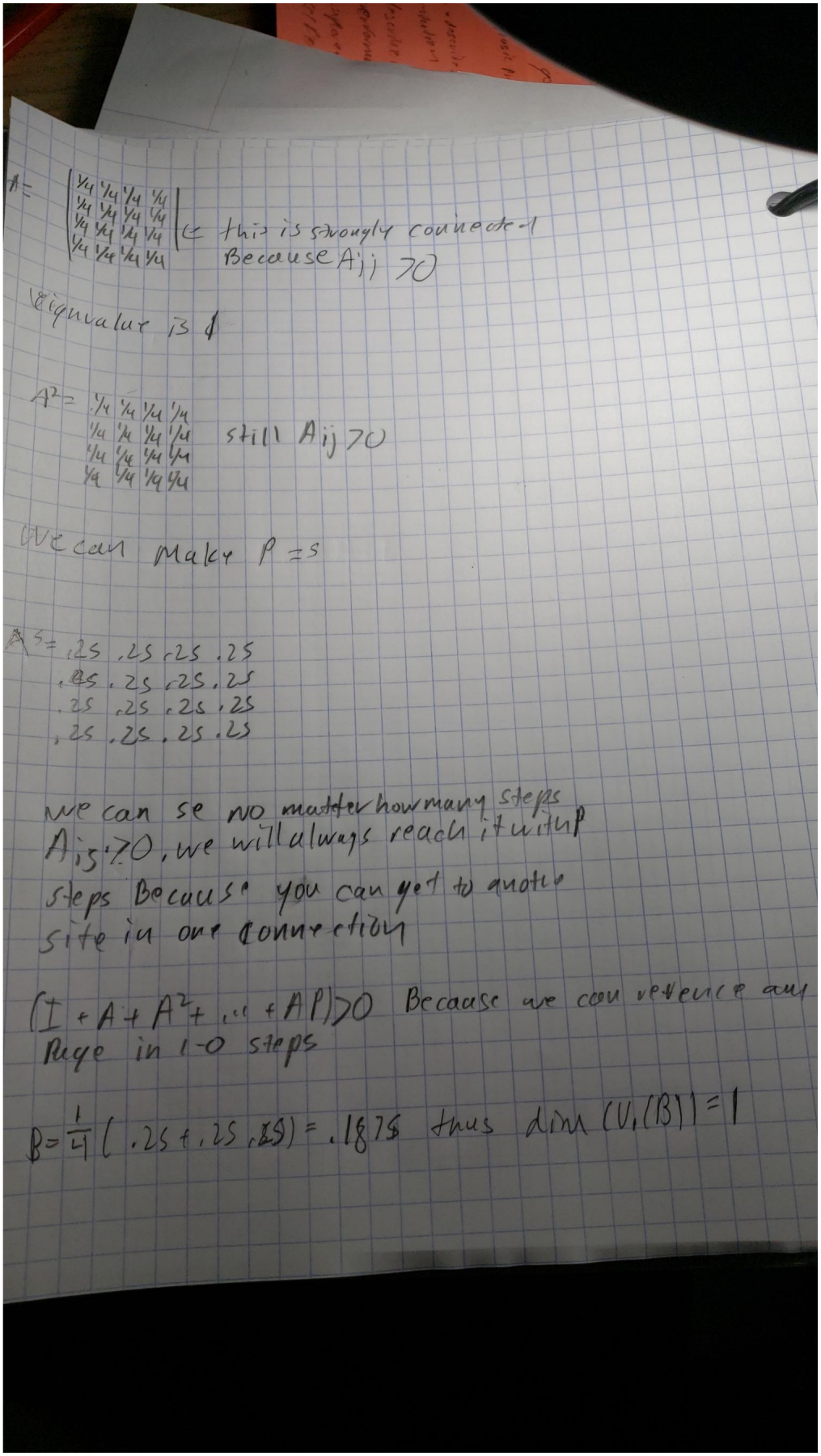


Figure 5: Answers to Exercise 10

- Exercise 11. Consider again the web in Figure 2.1, with the addition of a page 5 that links to page 3, where page 3 also links to page 5. Calculate the new ranking by finding the eigenvector of M (corresponding to $\lambda = 1$) that has positive components summing to one. Use m = 0.15.

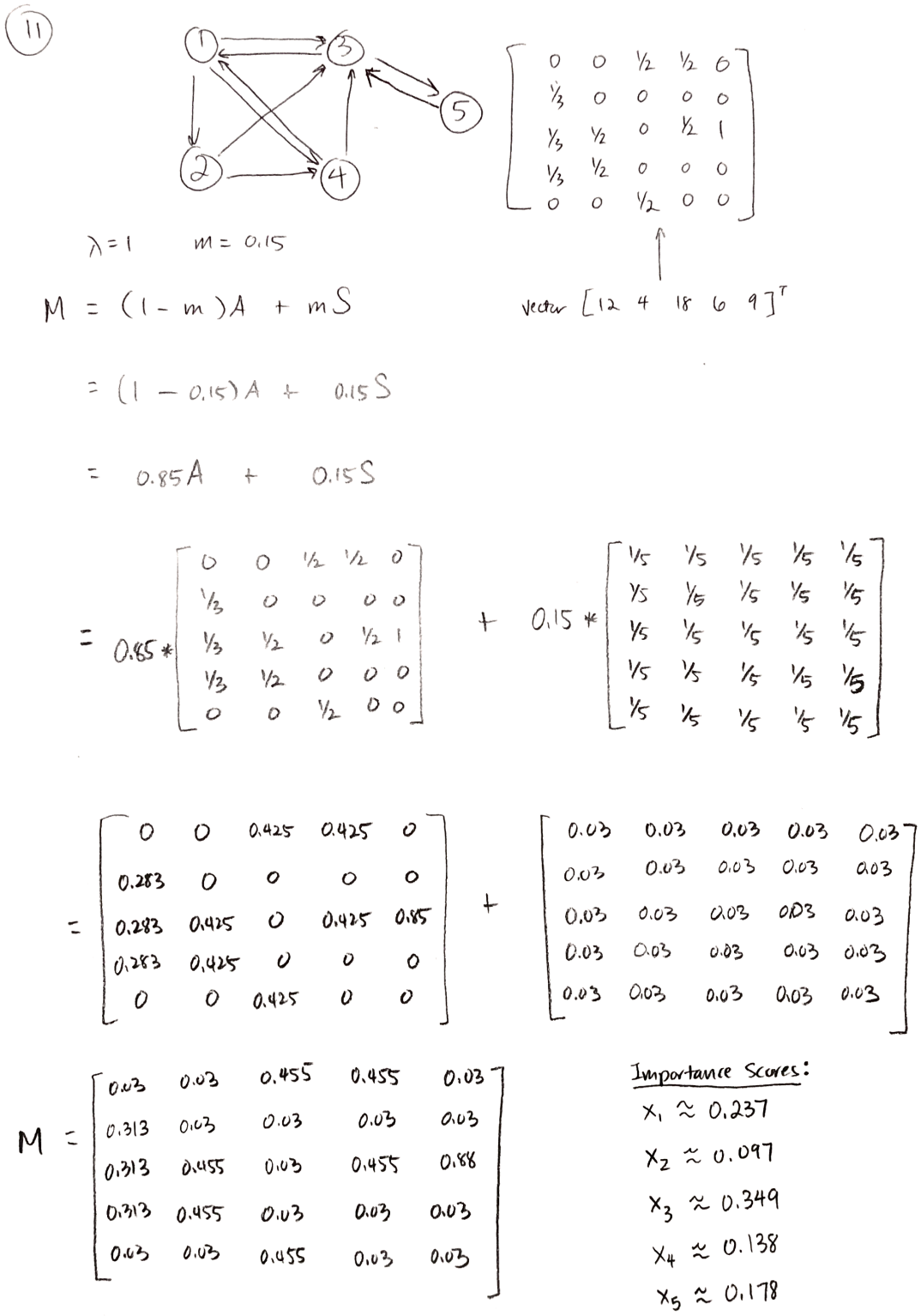
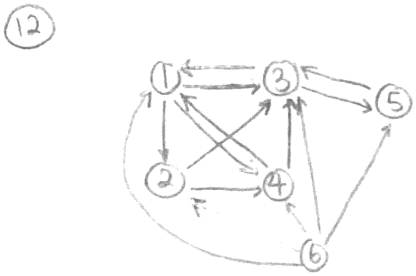


Figure 6: Answers to Exercise 11

- Exercise 12. Add a sixth page that links to every page of the web in the previous exercise, but to which no other page links. Rank the pages using A, then using M with m = 0.15, and compare the results.



$$A = \begin{bmatrix} 0 & 0 & 1/2 & 1/2 & 0 & 1/5 \\ 1/3 & 0 & 0 & 0 & 0 & 1/5 \\ 1/3 & 1/2 & 0 & 1/2 & 1 & 1/5 \\ 1/3 & 1/2 & 0 & 0 & 0 & 1/5 \\ 0 & 0 & 1/2 & 0 & 0 & 1/5 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$x_1 \approx 0.245$
$x_2 \approx 0.082$
$x_3 \approx 0.367$
$x_4 \approx 0.122$
$x_5 \approx 0.184$
$x_6 = 0$

$M = (1-m)A + mS \qquad \lambda = 1 \qquad m = 0.15$

$= 0.85A + 0.15S$

$$= 0.85 * \begin{bmatrix} 0 & 0 & 1/2 & 1/2 & 0 & 1/5 \\ 1/3 & 0 & 0 & 0 & 0 & 1/5 \\ 1/3 & 1/2 & 0 & 1/2 & 1 & 1/5 \\ 1/3 & 1/2 & 0 & 0 & 0 & 1/5 \\ 0 & 0 & 1/2 & 0 & 0 & 1/5 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} + 0.15 * \begin{bmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & 0.425 & 0.425 & 0 & 0.17 \\ 0.283 & 0 & 0 & 0 & 0 & 0.17 \\ 0.283 & 0.425 & 0 & 0.425 & 0.85 & 0.17 \\ 0.283 & 0.425 & 0 & 0 & 0 & 0.17 \\ 0 & 0 & 0.425 & 0 & 0 & 0.17 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 \end{bmatrix}$$

Figure 7: Answers to Exercise 12, p1

$$= \begin{bmatrix} 0.025 & 0.025 & 0.95 & 0.45 & 0.025 & 0.195 \\ 0.308 & 0.025 & 0.025 & 0.025 & 0.025 & 0.195 \\ 0.308 & 0.45 & 0.025 & 0.45 & 0.875 & 0.195 \\ 0.308 & 0.45 & 0.025 & 0.025 & 0.025 & 0.195 \\ 0.025 & 0.025 & 0.45 & 0.025 & 0.025 & 0.195 \\ 0.025 & 0.025 & 0.025 & 0.025 & 0.025 & 0.025 \end{bmatrix}$$

$x_1 \approx 0.231$
$x_2 \approx 0.095$
$x_3 \approx 0.340$
$x_4 \approx 0.135$
$x_5 \approx 0.174$
$x_6 \approx 0.025$

- The importance scores are fairly similar using either A or M, with the variance being very negligible (ranging from 0.010 to 0.027). The order of rank is the same for both A and M.

Figure 8: Answers to Exercise 12, p2

- Exercise 13. Construct a web consisting of two or more subwebs and determine the ranking given by formula (3.1).

– see exercise13.py file

- Exercise 14. For the web in Exercise 11, compute the values of $||M^k x_0 - q||_1$ and $\frac{||M^k x_0 - q||_1}{||M^{k1} x_0 - q||_1}$ for $k = 1, 5, 10, 50$, using an initial guess x_0 not too close to the actual eigenvector q (so that you can watch the convergence). Determine $c = max_{1 \le j \le n} \mid 1 - 2min_{1 \le i \le n} M_{ij} \mid$ and the absolute value of the second largest eigenvalue of M.

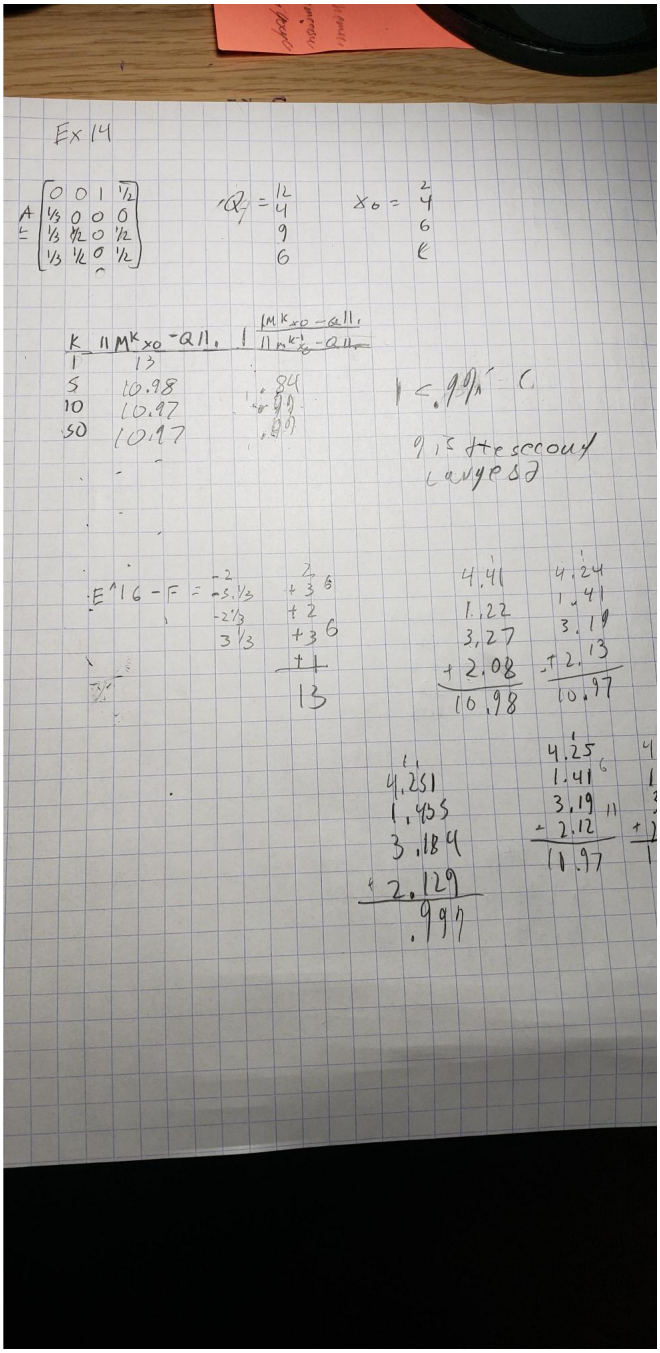


Figure 9: Answers to Exercise 14

- Exercise 15. To see why the second largest eigenvalue plays a role in bounding $\frac{||M^k x_0 - q||_1}{||M^{k1} x_0 - q||_1}$, consider an n n positive column-stochastic matrix M that is diagonalizable. Let x_0 be any

vector with non-negative components that sum to one. Since M is diagonalizable, we can create a basis of eigenvectors $q, v_1, ..., v_{n1}$, where q is the steady state vector, and then write $x_0 = aq + \Sigma_{k=1}^{n1} b_k v_k$. Determine $M^k x_0$, and then show that a = 1 and the sum of the components of each v_k must equal 0. Next apply Proposition 4 to prove that, except for the non-repeated eigenvalue $\lambda = 1$, the other eigenvalues are all strictly less than one in absolute value. Use this to evaluate $lim_{k \longrightarrow 1} \frac{||M^k x_0 - q||_1}{||M^{k1} x_0 - q||_1}$.

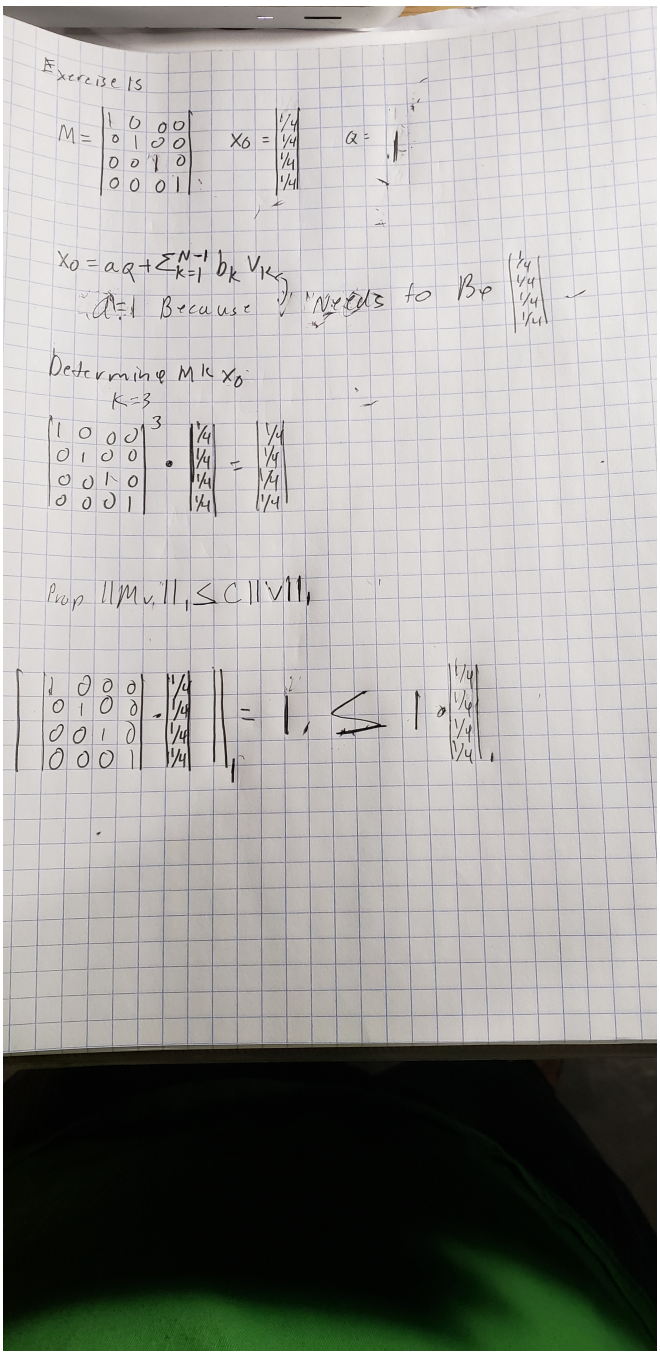


Figure 10: Answers to Exercise 15

- Exercise 16. Consider the link matrix

$$A = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} \\ 1 & \frac{1}{2} & 0 \end{bmatrix}$$

Show that $M = (1-m)A + mS(allS_{ij} = 1/3)$ is not diagonalizable for $0 \leq m < 1$.

– $M = (1-m)A + mS(allS_{ij} = 1/3)$ is still a stochastic matrix, as proved in exercise 8. If A is not diagnalizable, a linear combination of A and S: M is not either.

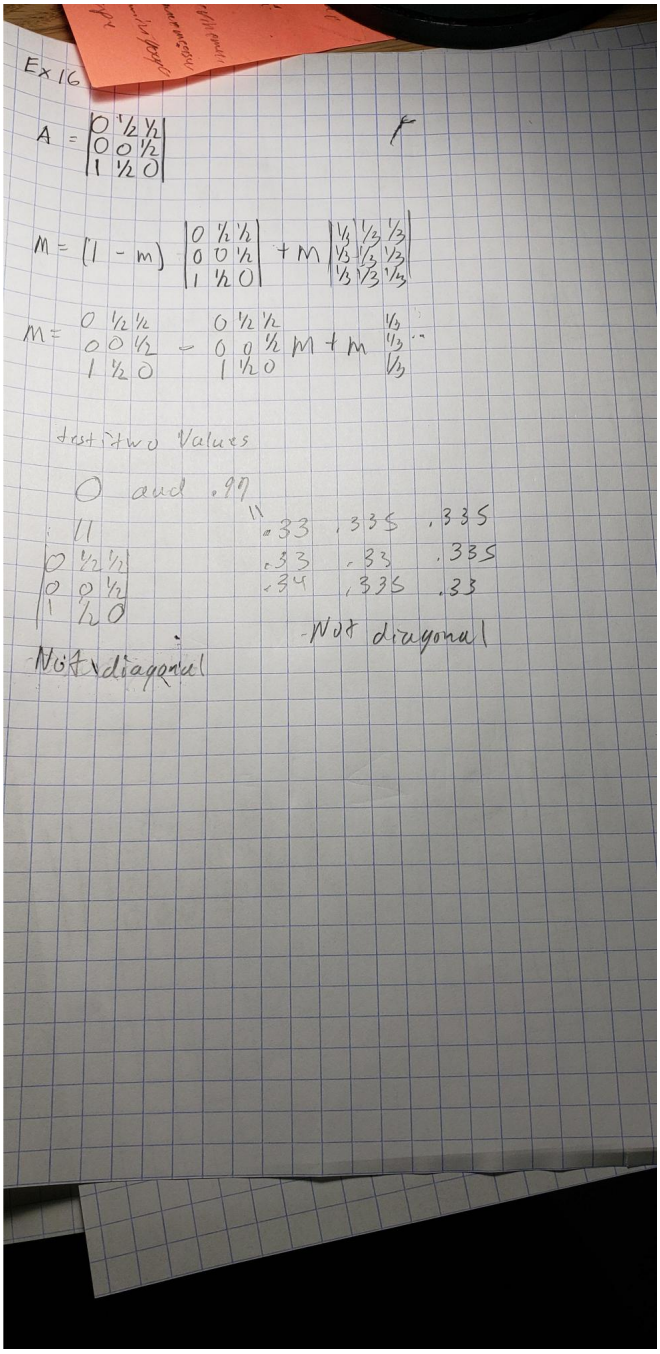


Figure 11: Answers to Exercise 16

- Exercise 17. How should the value of m be chosen? How does this choice affect the rankings and the computation time?
  - The value of m determines the importance score given to dangling nodes, $\frac{m}{n}$. By increasing the value of n, we increase the importance placed on dangling nodes relative to the linked portion of the web, resulting in a flatter distribution of importance scores. In the case of disjoint networks it likely makes sense to have a higher value of m, however for most purposes, a small, non-zero value (e.g. 0.15) is likely sufficient. Regarding performance, we hypothesize that performance would suffer when a high value of m is chosen for a a large well connected network and when a low value of m is chosen for a large disjoint network.