# Homework 3

These problems are due in-class on November 3rd.

## Problem 1 (5 pts)

For a simple, undirected graph, show that the sum of the eigenvalues of its adjacency matrix must be equal to 0.
(This is a quick problem.)

**Solution** The trace of the matrix is 0 because there are no self-loops, so the sum of the eigenvalues is too.

## Problem 2 (10 pts)

You've recently been hired as a senior researcher at Bingle!, a new scrappy search startup. They want to test out some algorithms for weakly preferential PageRank because they don't think that resetting "everywhere" from a dangling node is a realistic model. However, they only have a routine to compute a strongly preferential PageRank vector. That is, they have a program `pagerank(alpha,v)` that solves the system

$$(\boldsymbol{I} - \alpha\bar{\boldsymbol{P}} - \alpha\mathbf{v}\mathbf{d}^T)\mathbf{x} = (1-\alpha)\mathbf{v}$$

on their distributed MapReduce cluster for their current crawl of the web. The matrix $\bar{\boldsymbol{P}}$ is a column sub-stochastic matrix from a hand-tuned random walk process on their web graph, and $\mathbf{d} = \mathbf{e}^T - \mathbf{e}^T\bar{\boldsymbol{P}}$ is the dangling node vector
It's very time-consuming and expensive to change that code (and so, for this problem, you cannot). They've asked you to determine if there is another way to compute weakly preferential PageRank.

Thankfully, you've taken CS-59000 NMC at Purdue, and this is your chance to shine and earn your yearly bonus early. Show how to compute a weakly preferential PageRank vector by using their `pagerank(alpha,v)` routine and one additional routine `mult(x)` which computes $\bar{\boldsymbol{P}}\mathbf{x}$. Psst, the weakly preferential PageRank system is:

$$(\boldsymbol{I} - \alpha\bar{\boldsymbol{P}} - \alpha\mathbf{u}\mathbf{d}^T)\mathbf{y} = (1-\alpha)\mathbf{v}.$$

**Solution** The simplest way to solve the problem is to notice that we can convert $(\boldsymbol{I} - \alpha\bar{\boldsymbol{P}} - \alpha\mathbf{v}\mathbf{d}^T)\mathbf{x} = (1-\alpha)\mathbf{v}$ into the solution of a PseudoRank system by computing: $\gamma = \mathbf{d}^T\mathbf{x} = \mathbf{e}^T(\mathbf{x} - \boldsymbol{P}\mathbf{x})$, which involves a multiplication, a difference, and a summation. Thus, we find $\hat{\mathbf{x}} = \mathbf{x}/(1-\alpha+\alpha\gamma)$ satisfies

$$(\boldsymbol{I} - \bar{\boldsymbol{P}})\hat{\mathbf{x}} = \mathbf{v}.$$

We can do the same thing with the solution of $(\boldsymbol{I} - \alpha\bar{\boldsymbol{P}} - \alpha\mathbf{v}\mathbf{d}^T)\mathbf{w} = (1 - \alpha\mathbf{u}$ as well. This means we have all the ingredients to apply the Sherman-Morrison formula. However, we can do better. (In fact, this is usually true when you apply the Sherman-Morrison formula – there is almost always a better way, so use it for intuition and not algorithms.)

Notice that

$$(\boldsymbol{I} - \alpha\bar{\boldsymbol{P}} - \alpha\mathbf{u}\mathbf{d}^T)\mathbf{x} + \alpha\gamma\mathbf{u} - \alpha\gamma\mathbf{v} = (1-\alpha)\mathbf{v}.$$

Also, we have $\mathbf{u} = (1-\alpha)^{-1}(\boldsymbol{I} - \alpha\bar{\boldsymbol{P}} - \alpha\mathbf{u}\mathbf{d}^T)\mathbf{w}$. Consequently,

$$(\boldsymbol{I} - \alpha\bar{\boldsymbol{P}} - \alpha\mathbf{u}\mathbf{d}^T)(\mathbf{x} + \frac{\alpha\gamma}{1-\alpha}\mathbf{w}) = (1 - \alpha + \alpha\gamma)\mathbf{v}.$$

At this point, we are basically done, because we have a linear combination of $\mathbf{x}$ and $\mathbf{w}$ that is equal to a multiple of $\mathbf{v}$. Because the solution must sum to one, we just have to normalize this vector and we are done.

Hence, the algorithm is:

1. Compute `x␣=␣pagerank(alpha,v)`
2. Compute `w␣=␣pagerank(alpha,w)`
3. Compute `gamma␣=␣sum(x␣-␣mult(x))`
4. Set `y␣=␣x␣+␣alpha*gamma/(1- alpha)*w`
5. Update `y␣=␣y/sum(y)`

## Problem 3 (10 pts)

Google recently announced a change in how they process `rel=nofollow` links on the web. Please read about it here: http://www.mattcutts.com/blog/pagerank-sculpting/ Mathematically describe how this change affects how Google constructs the matrix $\boldsymbol{P}$ for PageRank.

**Solution**

Let $\boldsymbol{A}_1$ be the graph of links without no-follow and $\boldsymbol{A}_2$ be the graph of links with no-follow, and let $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ be the respective degree matrices. . Then the original PageRank formulation used $\boldsymbol{A}_1$ as the web-graph and normalized it as $\boldsymbol{D}_1^{-1}\boldsymbol{A}_1$ to get a random walk. The new formulation uses $(\boldsymbol{D}_1 + mD_2)^{-1}\boldsymbol{A}_1$ to normalize the walk.

## Problem 4 (10 pts)

Bingle! is back with another problem. One of their interns implemented a super-duper fast routine to compute a PseudoRank vector:

$$(\boldsymbol{I} - \alpha\bar{\boldsymbol{P}})\mathbf{y} = \mathbf{v}.$$

This intern claimed that they could replace all their PageRank computations using this routine if they just normalized $\mathbf{y}$, i.e. $\mathbf{x} = \mathbf{y}/\|\mathbf{y}\|$. Unfortunately, the intern was hired by the White House to help manage the government's data efforts and left without justifying this result. There is apparently no time left to handle Bingle!'s email's marked URGENT!

And so, they come to you. Can you solve the mystery of the normalization? To be precise, show exactly that $\|\mathbf{y}\|$ is the correct adjustment to $\mathbf{y}$.

**Solution** This is implicit in the first problem. The key is to show that $\mathbf{e}^T\mathbf{y} = \frac{1-\alpha\mathbf{d}^T\mathbf{y}}{1-\alpha}$ by using $\mathbf{e}^T\bar{\boldsymbol{P}}\mathbf{y} = \mathbf{e}^T\mathbf{y} - \mathbf{d}^T\mathbf{y}$ and $\mathbf{e}^T\mathbf{y} = \alpha\mathbf{e}^T\bar{\boldsymbol{P}}\mathbf{y} + 1$. Then show that $\mathbf{x} = \frac{1-\alpha}{1-\alpha\mathbf{d}^T\mathbf{y}}\mathbf{y}$ solves $(\boldsymbol{I} - \alpha\bar{\boldsymbol{P}} - \alpha\mathbf{v}\mathbf{d}^T)\mathbf{x} = (1-\alpha)\mathbf{v}$.

## Problem 5 (5 pts)

Tell me your preferred ranking of the following days for the in-class presentations:

- November 15th
- November 22nd

- November 29th
- December 1st

Rank each day from 1-4. There is no guarantee I'll be able to assign you to the day you'd like, but we'll see what we can do.

Due to scheduling details at the end of class, we will have to adjust the length of the presentations. We need two people to volunteer to split their presentation over two days (so we can do 2.5 presentations a day). If you'd be willing to split your presentation, please let me know.