PageRank

December 4, 2019

1 PageRank

In this notebook, you'll build on your knowledge of eigenvectors and eigenvalues by exploring the PageRank algorithm. The notebook is in two parts, the first is a worksheet to get you up to speed with how the algorithm works - here we will look at a micro-internet with fewer than 10 websites and see what it does and what can go wrong. The second is an assessment which will test your application of eigentheory to this problem by writing code and calculating the page rank of a large network representing a sub-section of the internet.

1.1 Part 1 - Worksheet

1.1.1 Introduction

PageRank (developed by Larry Page and Sergey Brin) revolutionized web search by generating a ranked list of web pages based on the underlying connectivity of the web. The PageRank algorithm is based on an ideal random web surfer who, when reaching a page, goes to the next page by clicking on a link. The surfer has equal probability of clicking any link on the page and, when reaching a page with no links, has equal probability of moving to any other page by typing in its URL. In addition, the surfer may occasionally choose to type in a random URL instead of following the links on a page. The PageRank is the ranked order of the pages from the most to the least probable page the surfer will be viewing.

```
In [ ]: # Before we begin, let's load the libraries.
        %pylab notebook
        import numpy as np
        import numpy.linalg as la
        from readonly.PageRankFunctions import *
        np.set_printoptions(suppress=True)
```

1.1.2 PageRank as a linear algebra problem

Let's imagine a micro-internet, with just 6 websites (**A**vocado, **B**ullseye, **C**atBabel, **D**romeda, **e**Tings, and **F**aceSpace). Each website links to some of the others, and this forms a network as shown,

The design principle of PageRank is that important websites will be linked to by important websites. This somewhat recursive principle will form the basis of our thinking.

Imagine we have 100 *Procrastinating Pat*s on our micro-internet, each viewing a single website at a time. Each minute the Pats follow a link on their website to another site on the micro-internet.