# Notes: Pagerank I

WE HAVE A GIGANTIC DATABASE FULL OF CUSTOMER BEHAVIOR INFORMATION.
EXCELLENT. WE CAN USE NON-LINEAR MATH AND DATA MINING TECHNOLOGY TO OPTIMIZE OUR RETAIL CHANNELS!
IF THAT'S THE SAME THING AS SPAM, WE'RE HAVING A GOOD MEETING HERE.
scottadams@aol.com
www.dilbert.com
© 2000 United Feature Syndicate, Inc.
11/13/00

## 1 Background

Pagerank is one of the primary tools used by search engines to return good search results. This set of notes covers some basic algorithms for computing the pagerank. I emphasize basic because there are literally hundreds of different algorithms that have been developed, and researchers are still developing new algorithms. These new, more advanced algorithms use concepts like distributed computing or more advanced math to improve their runtime, but they all return the same basic result: the top eigenvalue of a special graph called the *web graph*.

**References 1.** (recommended) Matt Cutts was formerly the head of Google's web spam team, and now runs the United States Digital Service (a recently created branch of the US government). Watch his video on "How Google Search Works", which discusses the importance of pagerank at a very high level.

`https://www.youtube.com/watch?v=KyCYyoGusqs`

**References 2.** Our primary text for this week is *Deeper Inside Pagerank* by Langville and Meyers. It is available on the github repo or at

`https://galton.uchicago.edu/~lekheng/meetings/mathofranking/ref/langville.pdf`

You will be responsible for sections 1, 2, 3, 5.1, 6.1, 6.2.

## 2 Definitions

**Problem 1.** The reading uses the following terms, but does not explicitly define them. Use wikipedia to find definitions for these terms.

NOTE: You are not responsible for exact, formal definitions of these terms. You just need to be able to use and understand them in context.

NOTE: It is common for data mining papers to not define common terms. When I'm reading papers, I'm constantly having to look up definitions on wikipedia for these terms, and the purpose of this problem is to get you practice with that. One of the challenges of wikipedia is that there's a lot of information about each of these terms that may or may not be relevant to this class. The main challenge of this problem is figuring out how to define these terms so that they help you understand the rest of this paper. I therefore recommend not looking up these terms in wikipedia until you encounter them in the reading; then provide definitions below that help you understand the reading.

1. markov chain

2. stationary vector

3. stochastic matrix

4. spectral radius

5. subdominant eigenvalue

**Problem 2.** Reproduce the definitions from the reading of the following terms below. (Your midterm quiz will be to reproduce these definitions exactly.)

1. irreducible matrix

2. primitive matrix

3. aperiodic markov chain

4. **P**

5. $\bar{\mathbf{P}}$

6. $\bar{\bar{\mathbf{P}}}$

7. $\boldsymbol{\pi}$

8. $\mathbf{v}$

## 3 “Trivial” Problems

The trivial problems in this section are designed to help you practice using the definitions above.

NOTE: Mathematicians define the word “trivial” to mean that the problem follows directly from the definitions without needing any major insights. Trivial problems can still take a long time to solve, however, because it takes a long time to understand the definitions. Muggles think that “trivial” problems should be “easy” to solve, but that’s only the case if you have a really strong understanding of the involved definitions. It’s okay if these trivial problems do not feel easy.

**Problem 3.** Give an example of:

1. a stochastic matrix

2. a non-stochastic matrix

3. an irreducible matrix

4. a reducible matrix

5. a primitive matrix

6. a non-primitive matrix

**Problem 4.** Answer the following questions.

1. Is the matrix $\mathbf{P}$ stochastic? irriducible? primitive?

2. Is the matrix $\bar{\mathbf{P}}$ stochastic? irriducible? primitive?

3. Is the matrix $\bar{\bar{\mathbf{P}}}$ stochastic? irriducible? primitive?

**Problem 5.** Either prove or give a counterexample to the following claims.

NOTE: Whenever I'm reading a data mining text, and I encounter a matrix, I always ask myself these sorts of questions about the matrix to ensure I understand what is going on. The purpose of these questions here is to get you into a similar habit.

HINT: Any claim which is true will have a "trivial" proof. When looking for counterexamples, use the identity and zero matrices as building-blocks.

1. $\text{rank}(\bar{\bar{\mathbf{P}}}) = 1$.

2. $\text{rank}(\bar{\bar{\mathbf{P}}}) = n$.

3. Let $\mathbf{x}$ be an eigenvector of $\mathbf{P}$ with eigenvalue $\lambda$. Then $\frac{1}{2}\mathbf{x}$ is also an eigenvector of $\mathbf{P}$ with eigenvalue $\frac{1}{2}\lambda$.

4. The smallest eigenvalue of $\bar{\bar{\mathbf{P}}}$ is exactly 0.

5. The largest eigenvalue of $\bar{\bar{\mathbf{P}}}$ is exactly 1.

6. The largest eigenvalue of $\mathbf{P}$ is exactly 1.

7. The largest eigenvector of $\mathbf{P}$ is simple. Recall that a simple eigenvalue has multiplicity 1. That is, there is exactly 1 eigenvector with the same eigenvalue.

8. The largest eigenvector of $\bar{\mathbf{P}}$ is simple.

9. The largest eigenvector of $\bar{\bar{\mathbf{P}}}$ is simple.