

Math 304 Handout: Linear algebra, graphs, and networks.

December 1, 2006

CONTENTS

1. Graphs and adjacency matrices.	1
2. Word search	3
3. Ranking the web pages.	4

1. GRAPHS AND ADJACENCY MATRICES.

Definition. A *graph* is a collection of vertices connected by edges. A *directed graph* is a graph all of whose edges have directions, usually indicated by arrows.

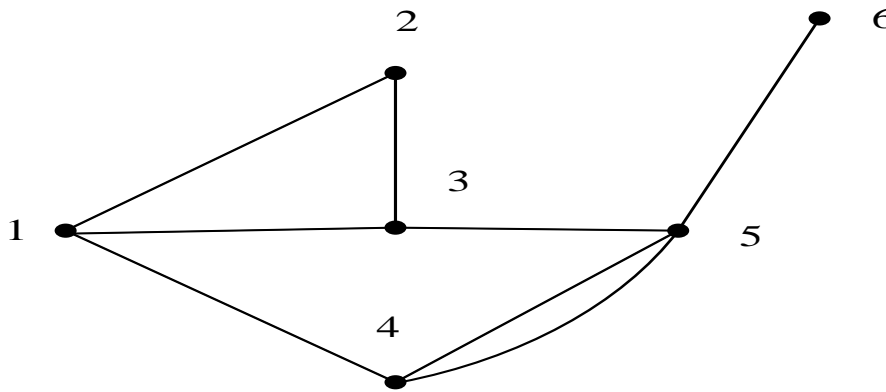


FIGURE 1. A graph.

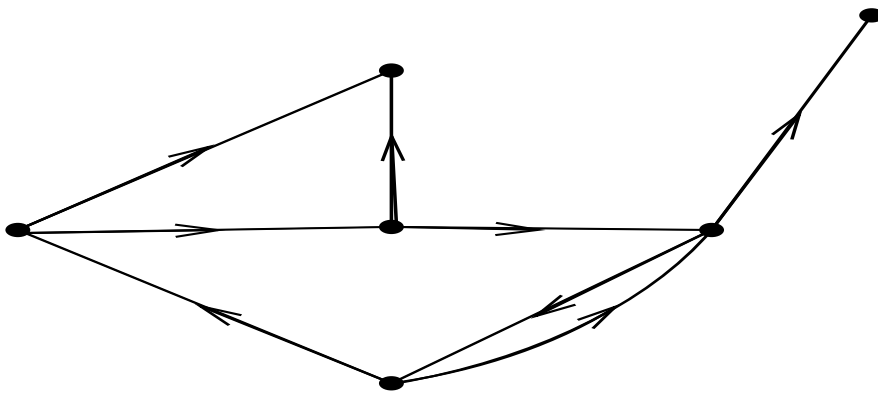


FIGURE 2. A directed graph.

Example.

- A DC electrical network is a directed graph.
- An AC electrical network is a graph.
- The Internet is a graph, with computers and servers as vertices and cables as edges.
- The World Wide Web is a directed graph, with web pages as vertices and links as edges.
- Towns connected by freeways form a graph.
- Squares connected by one-way streets form a directed graph.

Instead of drawing a graph, one can summarize it in a matrix. Label the vertices of a graph $1, 2, 3, \dots, n$. The **adjacency matrix** of a graph is the $n \times n$ matrix A such that a_{ij} is the number of edges connecting the j 'th and i 'th vertices. Similarly, for a directed graph the adjacency matrix is the $n \times n$ matrix A such that a_{ij} is the number of edges going from the j 'th to the i 'th vertex.

Example. The adjacency matrices for the graphs in Figures 1 and 2 are

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

and

$$B = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

For example, there is 1 edge connecting the 4th and the 1st vertex, 1 edge going from the 4th to the 1st vertex, and no edges going from the 1st to the 4th vertex. Note that the adjacency matrix of a graph is always symmetric.

Besides describing a graph as a bunch of numbers rather than as a picture, adjacency matrices have other uses. To find out which vertices can be reached from the 1st vertex in one step, we look at the entries of the 1st column of A . The non-zero entries are the 2nd, 3rd, and 4th, so those are the vertices that can be reached in one step. What about vertices that can be reached in two steps? It is not hard to see that these correspond to the entries in the column of A^2 .

$$A^2 = \begin{pmatrix} 3 & 1 & 1 & 0 & 3 & 0 \\ 1 & 2 & 1 & 1 & 1 & 0 \\ 1 & 1 & 3 & 3 & 0 & 1 \\ 0 & 1 & 3 & 5 & 0 & 2 \\ 3 & 1 & 0 & 0 & 6 & 0 \\ 0 & 0 & 1 & 2 & 0 & 1 \end{pmatrix}.$$

Thus in two steps, one can go from the 1st vertex back to itself in three ways (by visiting the 2nd, 3rd, or 4th vertices first), one can get to the 2nd vertex in one way (by passing through the 3rd vertex), and one still cannot get to the 6th vertex. In a large network, one can use the adjacency matrix to determine the distance between two vertices.

Exercise. Describe how to interpret the entries of the powers of the adjacency matrix of a *directed* graph.

Exercise. Explain the 6 entry in the A^2 matrix above.

2. WORD SEARCH

How do search engines work? The first issue is to find all the web pages; as far as I remember, the first efficient web crawler was Altavista in 1995. The second issue is to store, for all pages on the web, all the words that they contain. Again, an efficient way to do this is using a (very, very large) matrix. Each column corresponds to a web page, each row to a word, and each entry is 1 if a particular word occurs on a particular web page, and 0 otherwise.

Example. Suppose that we have 5 web pages, the search words are *linear*, *algebra*, *matrix*, *determinant*, *transformation*, and the occurrences of each word on each page are given by the following table:

	P1	P2	P3	P4	P5
linear	1	0	1	1	0
algebra	1	1	1	0	1
matrix	1	1	0	0	0
determinant	1	1	1	0	0
transformation	0	1	1	0	1

The corresponding matrix is

$$C = \begin{pmatrix} 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \end{pmatrix}$$

If we want to find out which web pages contain the terms “linear algebra”, we form the corresponding vector $\mathbf{x}_1 = (1, 1, 0, 0, 0)^T$ and calculate

$$C^T \mathbf{x}_1 = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 2 \\ 1 \\ 1 \end{pmatrix}.$$

Exercise. Why do we take C^T ?

We see that both words occur on the 1st and 3rd pages. Similarly, to find the pages containing the terms “determinant of a linear transformation”, we would calculate

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \\ 3 \\ 1 \\ 1 \end{pmatrix},$$

so that only the 3rd page contains all the terms.

3. RANKING THE WEB PAGES.

Once we know which web pages contain the desired words, how do we choose between these pages? An early approach was to return the pages that contain many occurrences of the search words. This idea is easily abused, since it is easy to put on a web page many repetitions of popular search words that have nothing to do with the page content.

A better approach is to use the graph structure of the Web. A page is “important” if many people think it is, and so there are many links to it. This idea is also easily defeated: a company can create lots of “dummy” sites to point to its main site, thereby giving an appearance that that site is popular. Instead, say that a site is important if there are many *important* sites that point to it. This seems like a circular definition, but one can make sense of it, and in 1999, Sergey Brin and Larry Page (following some ideas of Jon Kleinberg) based the Google search engine on it. Think of the Web as a giant Markov chain, where at every site a user randomly picks one of the links, goes on to the next site, again randomly picks a link, and continues in this manner. The matrix describing this Markov chain is almost the adjacency matrix of the directed graph, except that we want the matrix to be stochastic, and so normalize each column to have sum one. See the example below. The stationary vector of this matrix will tell us what proportion of time, on average, will a user spend at each site, which shows how popular or important that site is.

Example. Here is a network of 15 web pages and their links. The 5th site has many “naturally occurring” links pointing to it. The 14th site also has many links pointing to it, but they are “dummy” links since they all come from the 15th site. The stochastic matrix corresponding to this network is

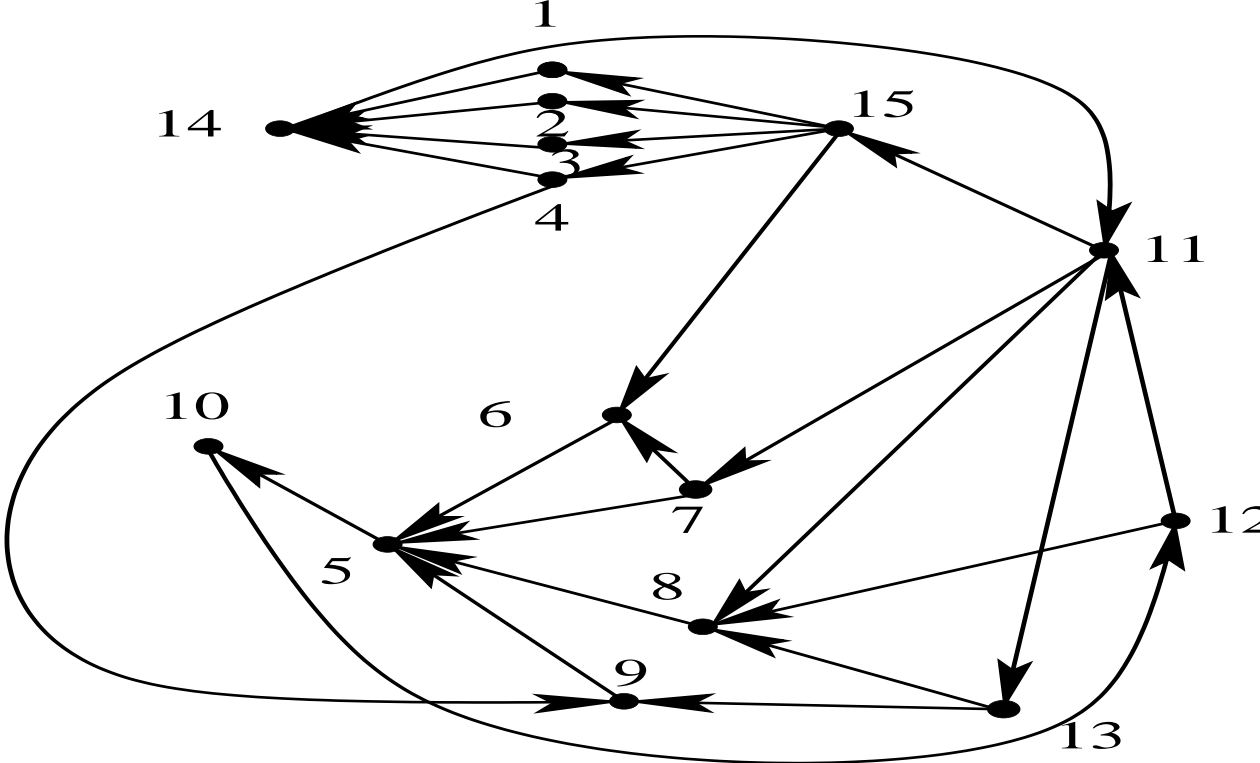


FIGURE 3. A large network.

$$D = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{2} & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Its eigenvalues are approximately

$$1, 0.19 \pm 0.85i, -0.19 \pm 0.70i, -0.87, -0.57, 0.44, 0, 0, 0, 0, 0, 0, 0$$

The eigenvector with eigenvalue 1 and positive entries that add up to 1 is approximately

$$(0.01, 0.01, 0.01, 0.01, 0.19, 0.02, 0.03, 0.14, 0.02, 0.19, 0.12, 0.19, 0.03, 0.02, 0.03)^T$$

Thus the ordering of the web pages is: 5, 10 (because 5 points to it), 12 (because 10 points to it), 8 and 11 (because 12 points to them), and the rest are unimportant.

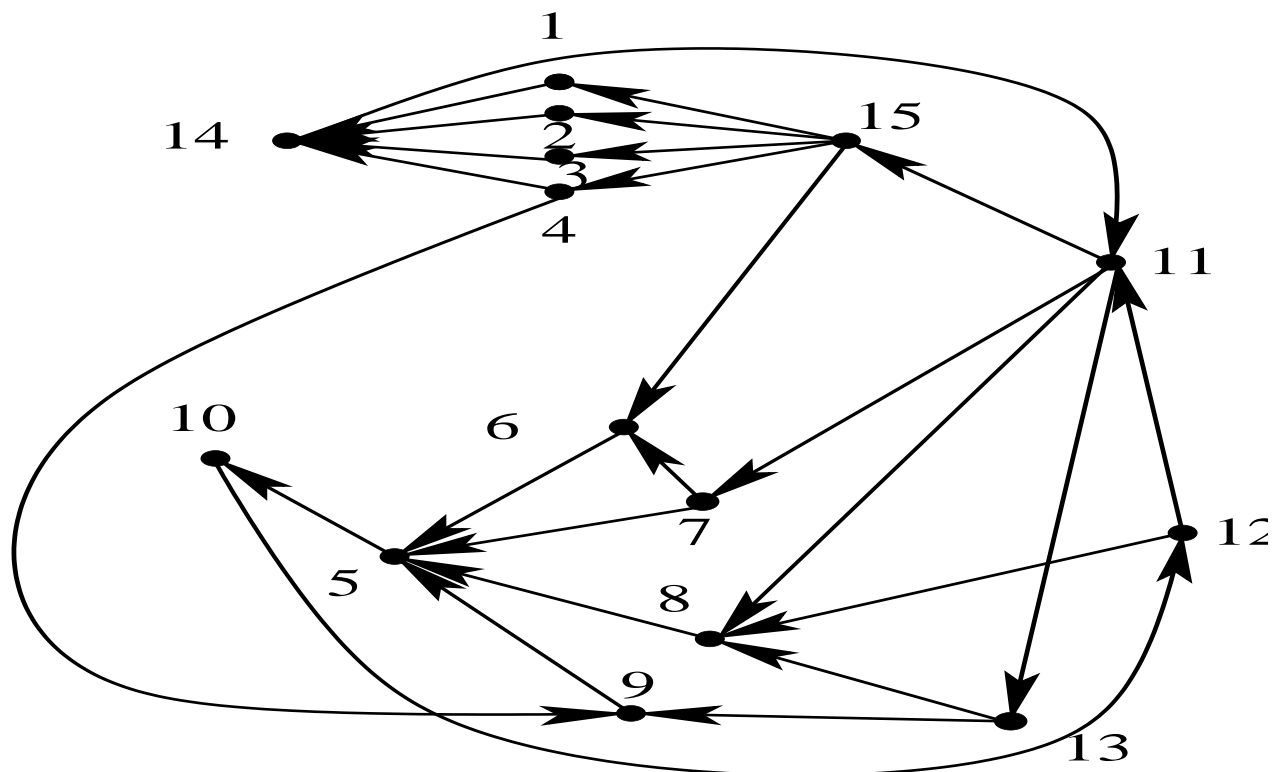


FIGURE 4. A large network.

Remark. This matrix does not have positive entries (it contains many zeros), so we cannot apply the theorem discussed in class to conclude that any initial distribution will lead to the stationary distribution. We *can* use Theorem 6.3.4 (since 1 is the dominant eigenvalue), or use the fact that the network is *transitive* (one can get from any vertex to any vertex) to conclude that some *power* of D has positive entries, which is enough. An alternative, actually implemented in Google, is to modify our model a little. We assume that at page, the user may follow any of the links (with equal probabilities), but also, with some small probability, jump to any other page. This is also a reasonable model, and mathematically it makes every entry of the matrix positive (if very small). The resulting eigenvalues and eigenvectors are not significantly different from the ones we calculated.

For more details, see the Wikipedia entry on “PageRank”.