CALIFORNIA STATE UNIVERSITY, LONG BEACH

EE 381 - Probability and Statistics with Applications to Computing

Laboratory Projects

Markov Chains

0. Introduction and Background Material

0.1. A simple example of a Markov chain

A machine can be in two possible states A and B. The state of the machine at time $(k+1)$ is determined by the state of the machine at the previous time step k:

- If the state at k is A, then the state at $(k+1)$ will be A with probability p_{11}
- If the state at k is A, then the state at $(k+1)$ will be B with probability p_{12}
- If the state at k is B, then the state at $(k+1)$ will be A with probability p_{21}
- If the state at k is B, then the state at $(k+1)$ will be B with probability p_{22}

These state transitions are shown schematically in Figure 1. Also, Figure 2 shows the same information on state transition probabilities in a more compact form.

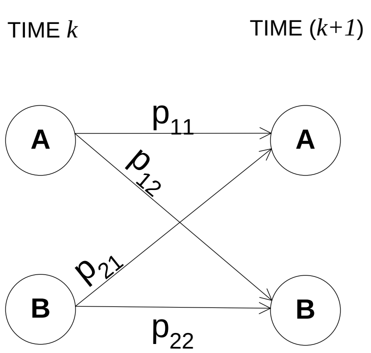


Figure 0.1

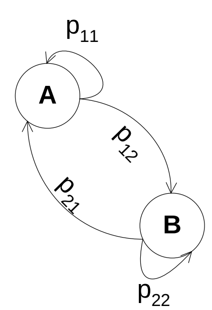


Figure 0.2

Clearly: $p_{11}+p_{12}=1$ and $p_{21}+p_{22}=1$.

For the current problem the transition probabilities have the following values:

$$p_{11} = 0.8 \quad ; \quad p_{12} = 0.2 \quad ; \quad p_{21} = 0.5 \quad ; \quad p_{22} = 0.5$$

The initial state is randomly distributed between state “A” and state “B”.
The probability of the initial state being “A” is given as: Prob (Initial State = A) = 0.4
The probability of the initial state being “B” is given as: Prob (Initial State = B) = 0.6

The graphs below show several simulation results for this two-state Markov chain. The chain is run for $n = 15$ time steps.

Figure 0.3 shows a single simulation run of the chain. The initial state of the chain is “A”. The graph shows the evolution of the chain for the next 15 steps. In this particular simulation the chain goes through the following sequence of states, after the initial state “A”:

$$\left[A \quad BBABBAABBBAAABA\right]$$

Figure 0.4 shows another single simulation run of the chain. In this case the initial state of the chain is “B”. The graph shows the evolution of the chain for the next 15 steps. In this particular simulation the chain goes through the following sequence of states, after the initial state “B”:

$$\left[B \quad AAABAAAAAAAAABA\right]$$

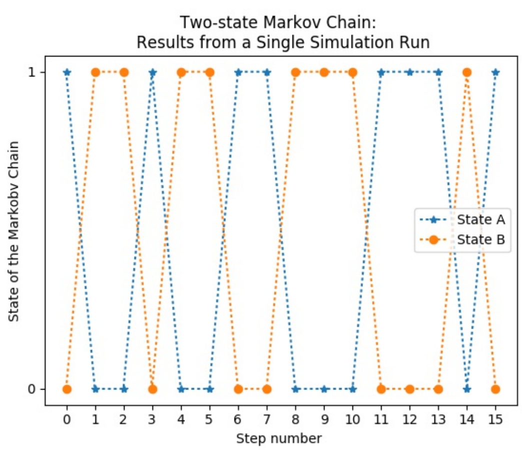


Figure 0.3

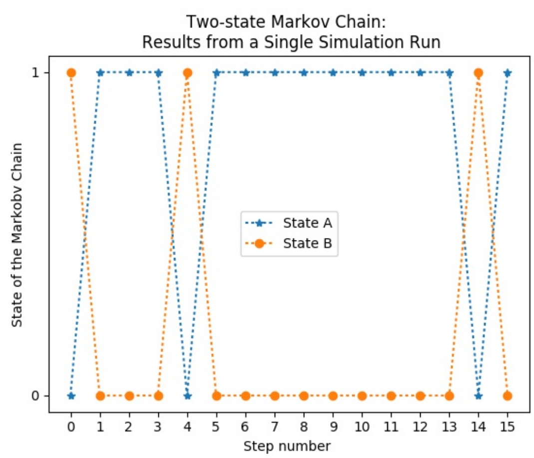


Figure 0.4

Figure 0.5 shows the combined results from N=10000 simulation runs of the chain. It is seen that the chain starts at state "A" 40% of the time, and at state "B" 60% of the time. The graph shows the evolution of the chain for the next 15 steps.

Figure 0.6 shows the results of the state evolution using the "State Transition Matrix" approach. The probabilities are *CALCULATED based on the transition matrix*. The probabilities they *ARE NOT derived as the result of N=10000 simulation runs*, as was done in Figure 0.5. It is noted that the results of Figures 0.5 and 0.6 are very similar, almost identical, confirming the fact that the state transition matrix approach can be used instead of a step-by-step simulation.

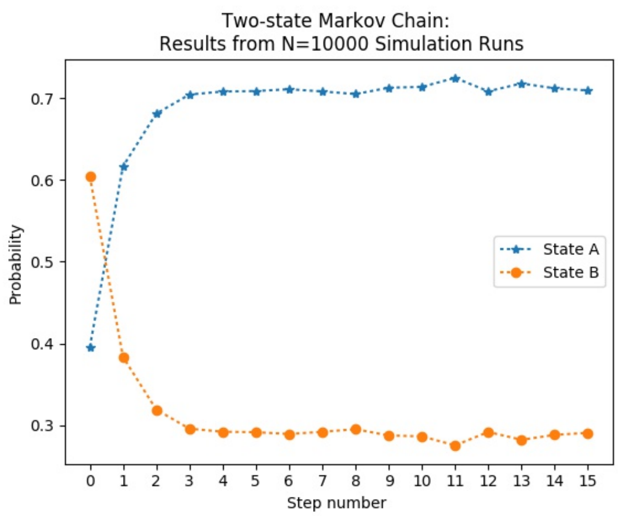


Figure 0.5

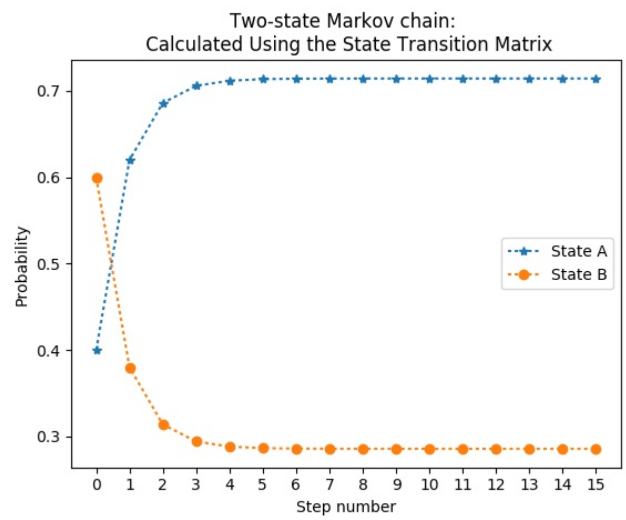


Figure 0.6

1. A three-state Markov Chain

Follow the previous code as a guide, and simulate the Markov chain defined in problem 11.1 of the AMS text by Grinstead & Snell, p. 406 (reference [1]).
The transition probabilities are given by the following matrix:

$$P = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}$$

1. The initial probability distribution of the state is given by: $\begin{bmatrix} R & N & S \end{bmatrix} = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}$
2. Run each experiment for $n = 15$ time steps. In order to obtain meaningful statistical data perform a total of $N = 10,000$ experiments.
3. After your experimental simulations are completed, use the State Transition Matrix approach and compare the results.

SUBMIT a report with the results and your code. You must follow the guidelines given in the syllabus regarding the structure of the report. Points will be taken off, if you do not follow the guidelines. The report should contain:

- Two plots showing two different single-simulation runs for $n = 15$ steps, similar to Figures 0.3 and 0.4.
- A plot with the experimental probabilities for the states R, N, S, similar to Figure 0.5
- A plot with the probabilities obtained through the state transition matrix approach for the three states R, N, S, similar to Figure 0.6
- The code in an Appendix
- Make sure that the plots are **properly labeled**

2. The Google PageRank Algorithm

This problem presents an introduction to the algorithm used for ranking web pages for searching purposes. The algorithm was developed by Google's founders Page and Brin with Motwani and Winograd, and was first published in 1998 (see reference [2]). The *PageRank* algorithm allowed *Google* to rise to the top of all web search engines within a matter of months after its implementation, outperforming the established search engines of the time such as *AltaVista* and *Yahoo*. The algorithm is based on the theory of Markov chains, utilizing the information on the number of links leading to an existing webpage.

The current problem uses a simplified web in order to show how the algorithm works (see reference [3]). The simplified web consists of 5 pages only, and it is shown schematically in Figure 2.1.

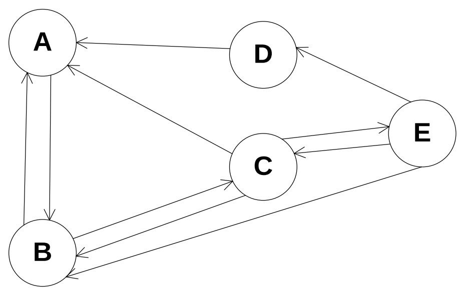


Figure 2.1: A five-page web

A Markov chain model of an *impartial surfer* visiting web pages is created as following:

- At time k the surfer is on page X, where $X \in \{A, B, C, D, E\}$.
- At time $(k+1)$ the surfer will move randomly to another page Y, which must be linked to X.
- The state S_k of the Markov chain at time k is defined as the page which the surfer is visiting at time k: $S_k \in \{A, B, C, D, E\}$.
- To create the Markov chain model, the algorithm assumes that all pages that can be reached from X have the same probability to be visited by the surfer. Thus:

$$\text{Prob}\left(S_{k+1} = Y \mid S_k = X\right) = \begin{cases} \dfrac{1}{\text{(Total number of links leaving page X)}} & , \quad \text{if } Y \text{ can be reached from } X \\ 0 & , \quad \text{if } Y \text{ cannot be reached from } X \end{cases}$$

- For example: $\text{Prob}\,(S_{k+1} = A \mid S_k = C) = \dfrac{1}{3}$; $\text{Prob}\,(S_{k+1} = E \mid S_k = D) = 0$; etc.

- Based on these probabilities the State Transition Matrix (P) of the Markov chain is constructed. The left eigenvector w of P corresponding to the eigenvalue $\lambda = 1$ (which is a *fixed vector* of the Markov chain) is computed from: $wP = w$
- The *PageRank* algorithm uses the vector w to rank the pages of the five-page web in terms of visiting importance.

To complete this problem follow the steps below:

1. Create the 5×5 State Transition Matrix P with the values of the transition probabilities
2. Assume that initially all pages are equally likely to be visited, i.e. use the initial state probability distribution vector: $v_1 = \begin{bmatrix} A & B & C & D & E \end{bmatrix} = \begin{bmatrix} \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \end{bmatrix}$
3. Calculate the probability vectors for $n = 1, 2, \ldots 20$ steps, **using the State Transition matrix only.** Note that *you ARE NOT asked to run the complete simulation* of the chain. You are only asked to use the State Transition matrix approach to calculate the probability vectors for n steps.
4. Rank the pages $\{A, B, C, D, E\}$ in order of importance based on the results from the previous step
5. Create a plot of the calculated state probabilities for each of the five states $\{A, B, C, D, E\}$ vs. the number of steps for $n = 1, 2, \ldots 20$. This should be similar to the plot in Figure 0.6
6. Assume that the surfer always starts at his/her home page, which for the purposes of this problem is page E. Repeat steps (2)-(5) with the new state probability distribution vector: $v_2 = \begin{bmatrix} A & B & C & D & E \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \end{bmatrix}$

SUBMIT a report with the results and your code. You must follow the guidelines given in the syllabus regarding the structure of the report. Points will be taken off, if you do not follow the guidelines. The report should contain:

- The 5×5 State Transition Matrix P
- For the initial state probability distribution vector v_1:
  - ❖ Submit the page ranking calculated in Step 4. *Use Table 1 for your answer*. Note: You will need to replicate the table in your Word file, in order to provide the answer in your report. Points will be taken off if you do not use the table.
  - ❖ Submit the plot created in Step 5.
- For the initial state probability distribution vector v_2
  - ❖ Submit the page ranking calculated in Step 4. *Use Table 1 for your answer*. Note: You will need to replicate the table in your Word file, in order to provide the answer in your report. Points will be taken off if you do not use the table.
  - ❖ Submit the plot created in Step 5.
- The code in an Appendix.
- Make sure the plots are **properly labeled**

Initial probability vector: v_1		
Rank	Page	Probability
1		
2		
3		
4		
5		

Initial probability vector: v_2		
Rank	Page	Probability
1		
2		
3		
4		
5		

Table 1.

References

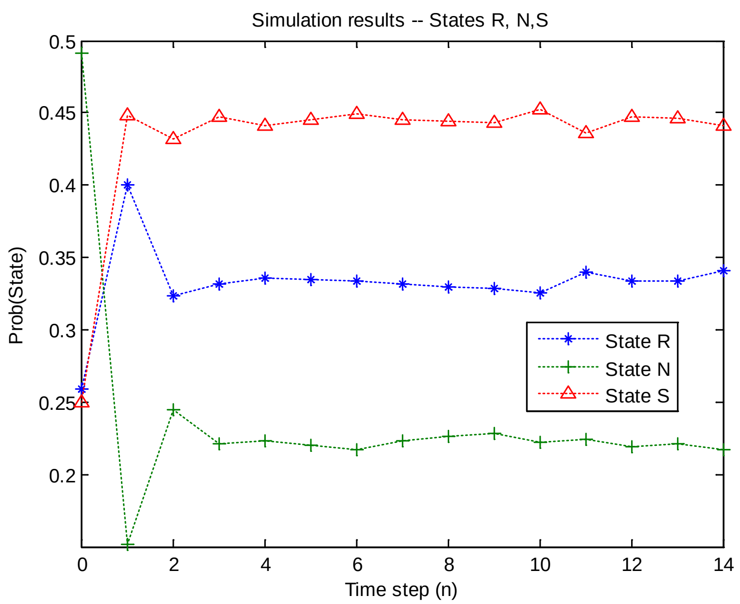
[1] "*Introduction to Probability*", C.M. Grinstead and J.L Snell, American Mathematical Society, electronic version, 2003, 2nd edition

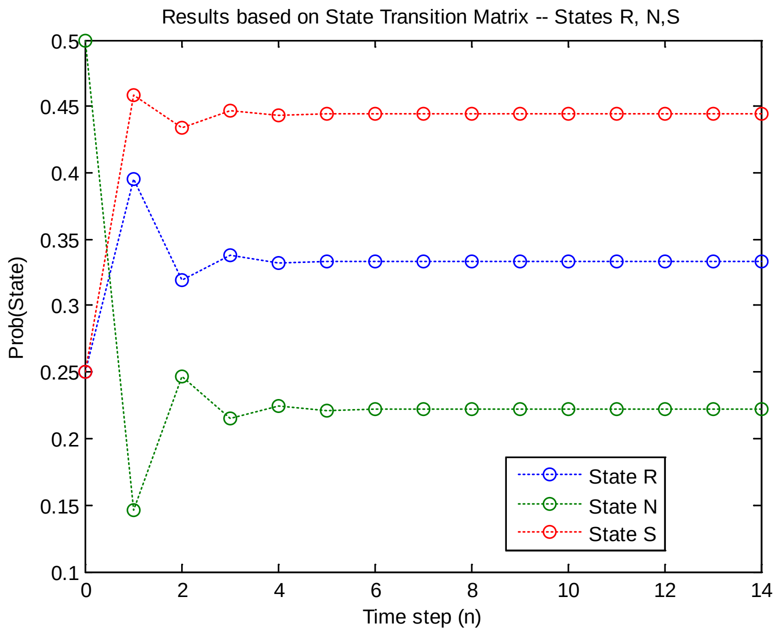
[2] L. Page, S. Brin, R. Motwani, and T. Winograd. "*The PageRank Citation Ranking: Bringing Order to the Web*", Technical Report, Stanford University, 1998.

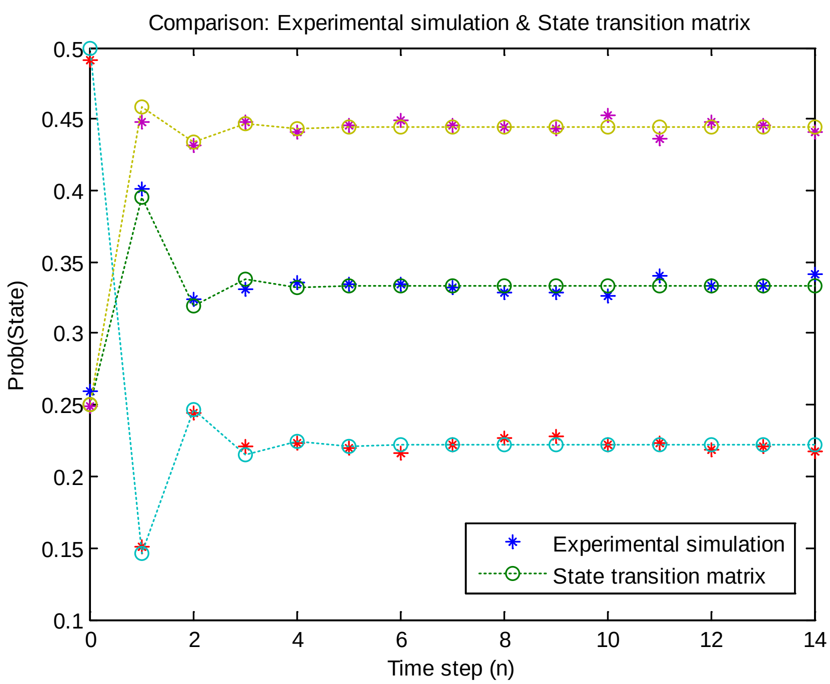
[3] "*Mathematics & Technology*", C. Rousseau & Y. Saint-Aubin, Springer, 2008.

```
%% PROBLEM 1 - THREE STATE MARKOV CHAIN
N=1e4;        % Number of experiments
n=15;         % Number of transitions to be computed
X=char(zeros(n,N)); % Each column of X represents one of the N experiments
M=zeros(n,3);       % M contains the experimental probabilities for states A & B
S=char(n,1);         % Initialize state array
%
p11=1/3; p12=1/3; p13=1/3;      % Transition probabilities
p21=1/2; p22=0; p23=1/2;
p31=1/4; p32=1/4; p33=1/2;
%
d01=1/4; d02=1/2; d03=1/4;  % Probability distribution for initial state
%
for j=1:N
    r0=rand();       % Calculate initial state
    if r0<=d01, s0='R';
    elseif (r0>d01 && r0<=d01+d02), s0='N';
    elseif r0>d01+d02; s0='S';
    end;
    S(1)=s0;            % Initial state
%
for k=1:n-1
    s=S(k);
    r=rand();
    if s=='R'
        if r<=p11, S(k+1)='R';
        elseif (r>p11 && r<=p11+p12), S(k+1)='N';
        elseif r>p11+p12, S(k+1)='S';
        end
    elseif s=='N'
        if r<=p21, S(k+1)='R';
        elseif (r>p21 && r<=p21+p22), S(k+1)='N';
        elseif r>p21+p22, S(k+1)='S';
        end
    elseif s=='S'
        if r<=p31, S(k+1)='R';
        elseif (r>p31 && r<=p31+p32), S(k+1)='N';
        elseif r>p31+p32, S(k+1)='S';
        end
    end;
end;
    X(:,j)=S;
end;
%
% Calculate the experimental probabilities for states A & B (matrix M)
for j=1:n
    x=X(j,:);
    ma=length(find(x=='R'));
    mb=length(find(x=='N'));
    mc=length(find(x=='S'));
    M(j,:)=[ma  mb mc]/N;
end;
%
```

```
nv=0:n-1;
figure(1);
plot(nv, M(:,1),'*:',nv, M(:,2),'+:',nv, M(:,3),'^:' );
title('Simulation results -- States R, N,S');
xlabel('Time step (n)');
ylabel('Prob(State)');
legend('State R', 'State N', 'State S')
%
% Use the state transition matrix approach and compare results
P=[p11 p12 p13; p21 p22 p23; p32 p32 p33];   % State transition matrix
y0=[1/4  1/2  1/4];      % Probability distribution of initial state
Y=zeros(n,3);
Y(1,:)=y0;
for k=1:n-1
    Y(k+1,:)=Y(k,:)*P;
end
%
figure(2); plot(nv,Y(:,1),'o:',nv,Y(:,2),'o:',nv,Y(:,3),'o:');
title('Results based on State Transition Matrix -- States R, N,S');
xlabel('Time step (n)')
ylabel('Prob(State)');
legend('State R', 'State N', 'State S')
%
% Compare the two approaches
figure(3);  plot(nv, M(:,1),'*',nv,Y(:,1),'o:',nv, M(:,2),'*',nv,Y(:,2),'o:',nv,
M(:,3),'*',nv,Y(:,3),'o:');
title('Comparison: Experimental simulation & State transition matrix');
legend('Experimental simulation','State transition matrix');
xlabel('Time step (n)')
ylabel('Prob(State)');
```




Results based on State Transition Matrix -- States R, N,S
Prob(State)
Time step (n)
State R
State N
State S
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0
2
4
6
8
10
12
14

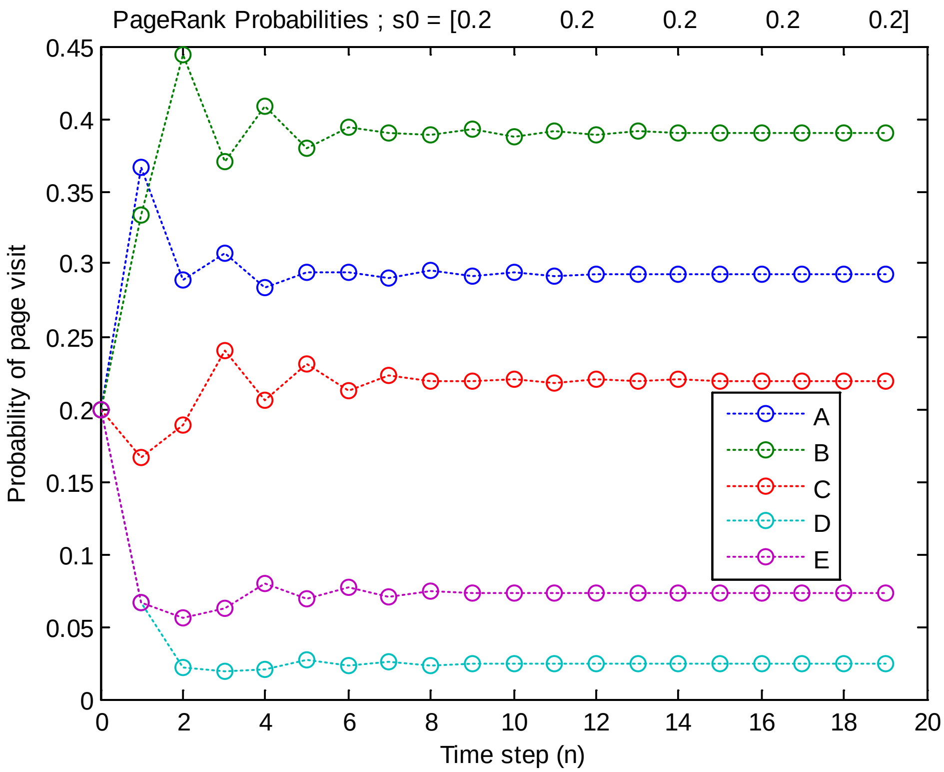

Comparison: Experimental simulation & State transition matrix
Prob(State)
Time step (n)
Experimental simulation
State transition matrix
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0
2
4
6
8
10
12
14

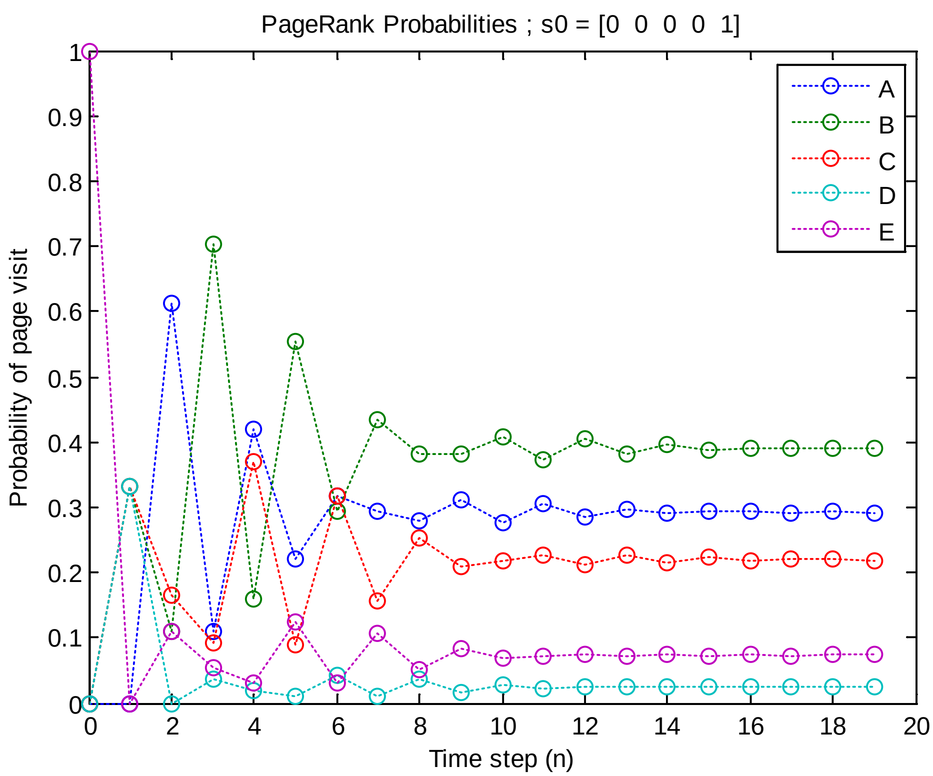
```
%% PROBLEM 2 ===== Google and the Page rank algorithm
% % C Rousseau & Y Saint-Aubin, "Mathematics & Technology", p.266
% % Springer, 2008
% % Five nodes A, B, C, D, E representing web pages
% % Use the state transition matrix approach only
%
N=1e4;         % Number of experiments
n=20;          % Number of transitions to be computed
X=char(zeros(n,N)); % Each column of X represents one of the N experiments
M=zeros(n,5);      % M contains the experimental probabilities for states A & B
S=char(n,1);       % Initialize state array
%
P=[0 1 0 0 0 ; 1/2 0 1/2 0 0 ; 1/3  1/3  0 0 1/3 ;  ...
    1 0 0 0 0 ; 0 1/3 1/3 1/3 0 ];
%
v=[1/5 1/5 1/5 1/5 1/5; 0 0 0 0 1];

for j=1:2
s0=v(j,:);  % Initial state probability distribution
Y=zeros(n,5);
Y(1,:)=s0;
for k=1:n-1
    Y(k+1,:)=Y(k,:)*P;
end

nv=0:n-1;

figure;
plot(nv,Y,'o:' );
title(['PageRank Probabilities', ' ; s0 = [',num2str(s0), ']']);
xlabel('Time step (n)');
ylabel('Probability of page visit')
legend('A','B','C','D','E');
%
end
```

```
[V,D]=eig(P')
w=V(:,1)/sum(V(:,1));   w=w'
w=[  0.2927    0.3902    0.2195    0.0244    0.0732]
```

Initial probability vector: v_1		
Rank	Page	Probability
1	B	0.3902
2	A	0.2927
3	C	0.2195
4	E	0.0732
5	D	0.0244

Initial probability vector: v_2		
Rank	Page	Probability
1	B	0.3902
2	A	0.2927
3	C	0.2195
4	E	0.0732
5	D	0.0244


PageRank Probabilities ; s0 = [0.2 0.2 0.2 0.2 0.2]
Probability of page visit
Time step (n)
A
B
C
D
E


PageRank Probabilities ; s0 = [0 0 0 0 1]
Probability of page visit
Time step (n)
A
B
C
D
E