Graphs 2: Link Analysis

300958 Social Web Analytics

WESTERN SYDNEY
UNIVERSITY

School of Computing, Engineering and Mathematics

Week 8

Outline

Motivation

Similar to centrality measures, link analysis methods allow us to identify the popularity of a vertex, based on the structure of the whole graph.

Google uses the link analysis method *PageRank* to compute the popularity score for each Web page. This score is provided as an indicator as to how high the Web page should be ranked in search results.

Outline

Random Walk

A random walk is a traversal of some space, where we take *n* steps (some number of steps), and each step is decided randomly.

Examples of random walks in 1, 2, and 3 dimensional space can be seen at:
`http://en.wikipedia.org/wiki/Random_walk`

Random Walk on a Graph

Random walks can be taken on graphs, where each step moves us to a vertex and the edges of each vertex provide a path between vertices.

Two random walks from v_1 of length 3

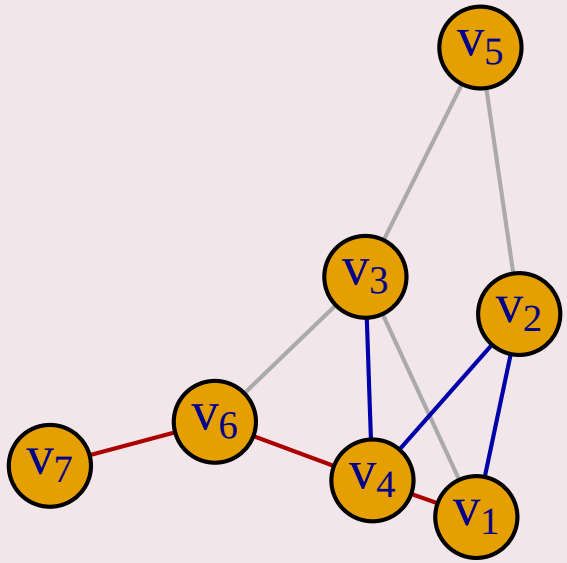


Figure: Two random walks of length 3 starting from v_1 shown in red and blue.

Random Walk on a Graph and Probabilities

By taking a random walk of length 1, we choose an edge at random that is connected to the current vertex and follow it. The probability of following a particular edge is equal to $1/\text{degree}(v)$, starting at vertex v.

Probability of arriving at vertex.

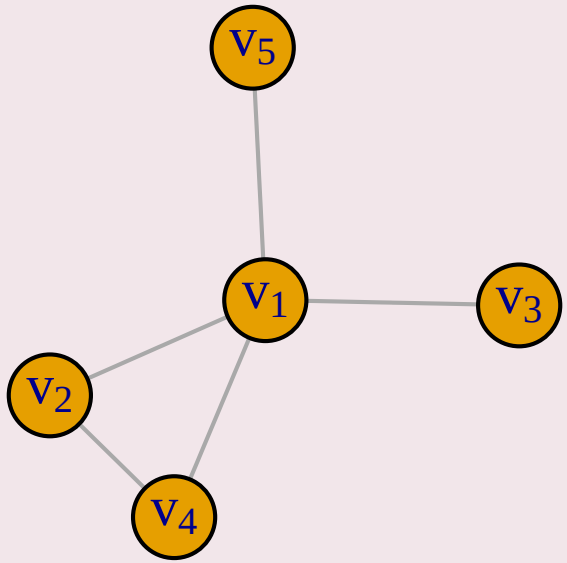


Figure: After a random walk of length 1 beginning at v_1, the probability of arriving at v_2 is $1/4$.

Random walk probability

Problem

After a random walk of length 1:

- What is the probability of arriving at v_3, when beginning at v_1?
- What is the probability of arriving at v_2, when beginning at v_4?
- What is the probability of arriving at v_5, when beginning at v_3?

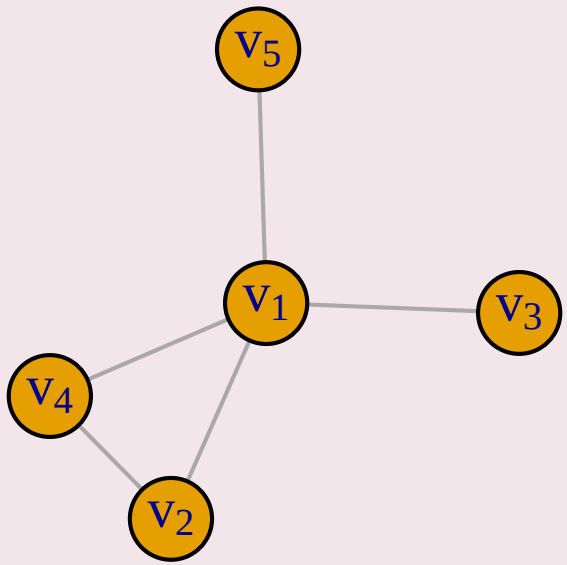


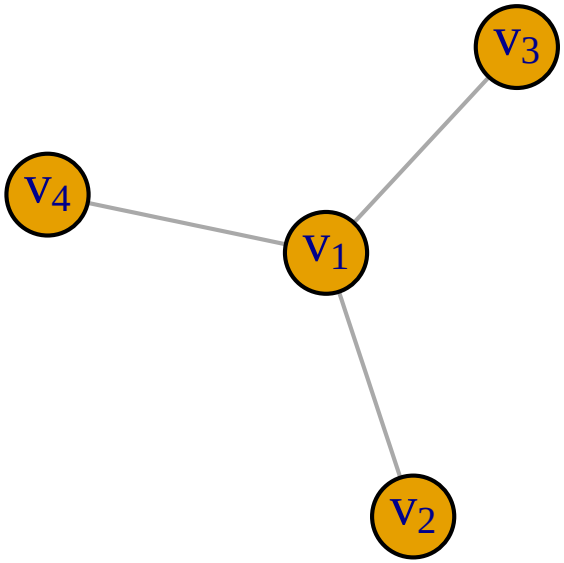
Figure: Random walk of length 1 problem.

State distribution

The state of a random walk is the vertex in which the walk has ended on. After a random walk of length n, there is a chance of being in multiple states (arriving at multiple vertices), where each state has a probability.

The collection of probabilities forms a distribution over the set of states.

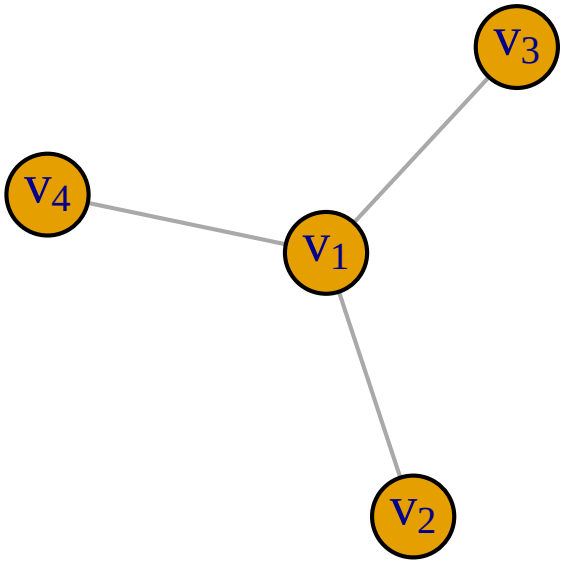
State distribution example



A random walk of length n, starting at v_1 gives the state distribution:

State	v_1	v_2	v_3	v_4
$\vec{p}_1$	0	1/3	1/3	1/3

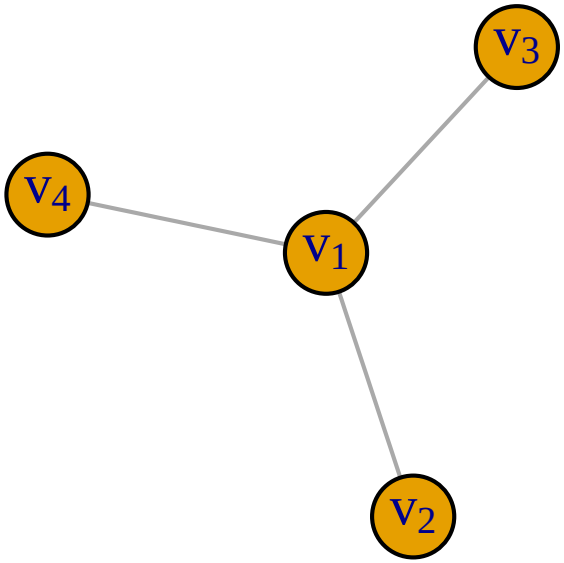
State distribution example



A random walk of length n, starting at v_1 gives the state distribution:

State	v_1	v_2	v_3	v_4
$\vec{p}_1$	0	1/3	1/3	1/3
$\vec{p}_2$	1	0	0	0

State distribution example



A random walk of length n, starting at v_1 gives the state distribution:

State	v_1	v_2	v_3	v_4
$\vec{p}_1$	0	1/3	1/3	1/3
$\vec{p}_2$	1	0	0	0
$\vec{p}_3$	0	1/3	1/3	1/3

State distribution

Problem

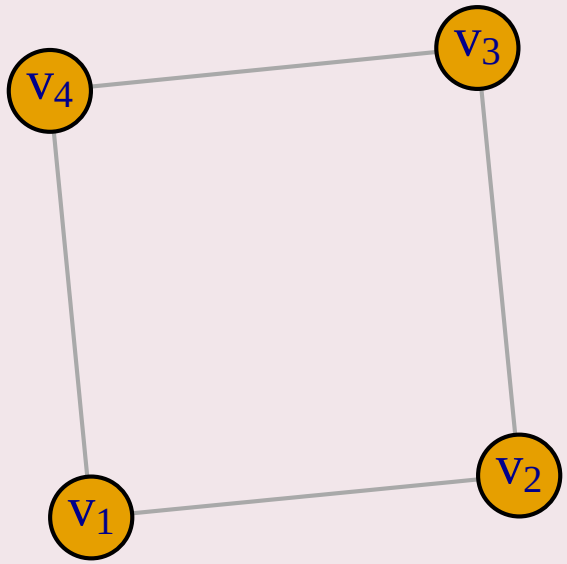


Figure: State distribution problem.

Compute the state distribution after a length $n = 1, 2, 3$ random walk, when starting at v_1.

Probability Transition

We find that the state distribution of a length n random walk can be computed using the graph and the knowledge of the previous state distribution (for a random walk of length $n-1$).

Therefore the state distribution is a Markov Chain (the current state n depends only on the previous state $n-1$, not on any other states before it).

We can compute the probability of a given state of a Markov Chain using its transition probability matrix T, where

$$\vec{p}_n = T\vec{p}_{n-1}$$

where $\vec{p}_n$ is the state distribution after a length n random walk.

Transition Probability Matrix

The transition probability matrix T of graph G provides us with the probability of moving to a state, given the probability of the current state.

T is a matrix with elements $t_{i,j}$ (ith row, jth column), where:

$$t_{i,j} = \begin{cases} 1/\text{degree}(v_j) & \text{if } e_{i,j} \in E \\ 0 & \text{otherwise} \end{cases}$$

All of the columns of T should sum to 1.

Transition Probability Matrix

Example

Write the transition probability matrix for the following graph:

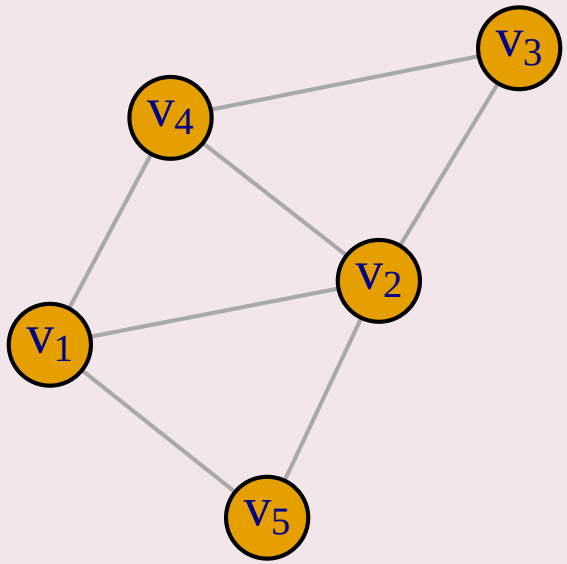


Figure: Transition probability matrix example.

Transition Probability Matrix

Problem

Write the transition probability matrix for the following graph:

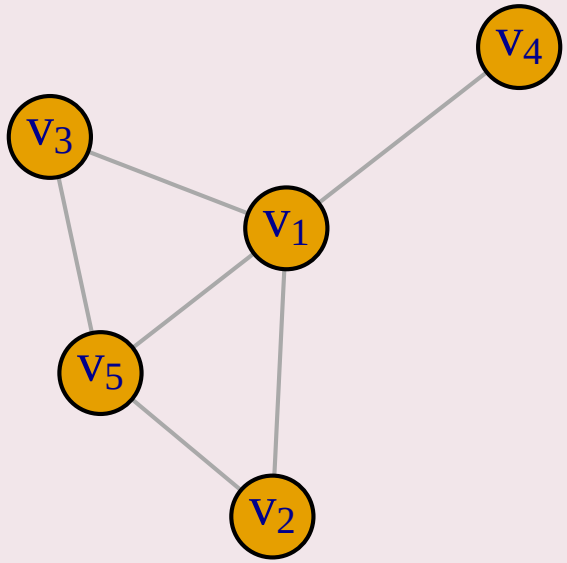
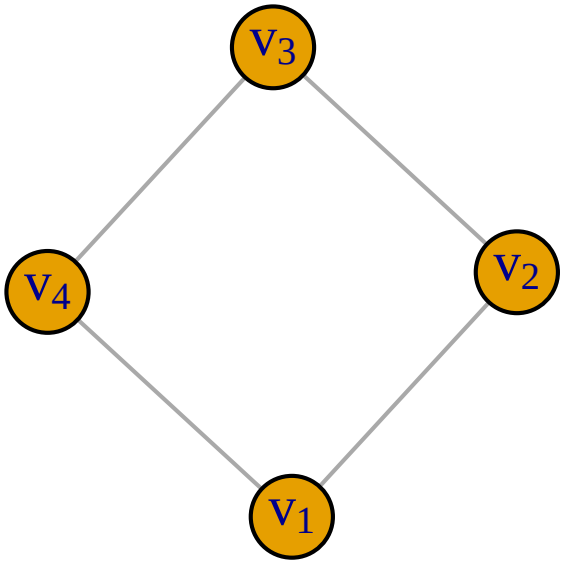


Figure: Transition probability matrix problem.

One Step Transition

Given any state distribution, we can take a random step by multiplying the state with the transition probability matrix. Let's examine the previous example.



We begin at v_1, so the initial state distribution is:

$$\vec{p} = [1\,0\,0\,0]$$

Let's find the state distribution after $n = 1, 2$ and 3 steps.

Infinite Random Walk on a Graph

The probability of arriving at vertex v after a random walk is a good measure of the vertex's popularity or influence in the graph.

We are able to compute the state distribution $\vec{p}_n$ using a random walk, but we have to provide an initial state $\vec{p}_0$, and the length of the random walk n. The probability will change depending on n and $\vec{p}_0$.

To avoid the problem of having to choose n and $\vec{p}$, we can instead examine the probability of arriving at vertex v after an infinite number of steps in a random walk. Using $n = \infty$, the state distribution $\vec{p}_\infty$ becomes independent of the initial state $\vec{p}_0$.

Stationary Distribution

The state distribution of a random walk on a graph converges, therefore:

$$\vec{p}_\infty = T\vec{p}_\infty$$

This implies that the state distribution at $n = \infty$ is the stationary distribution.

The stationary distribution $\vec{p}$ of graph G is defined as:

$$\vec{p} = T\vec{p}$$

where T is the graph transition probability matrix.

The stationary distribution is the probability distribution over the set of vertices, where taking a random step on the graph does not change the distribution (it is stationary).

Stationary Distribution Example

Example

Find the stationary distribution of the following graph:

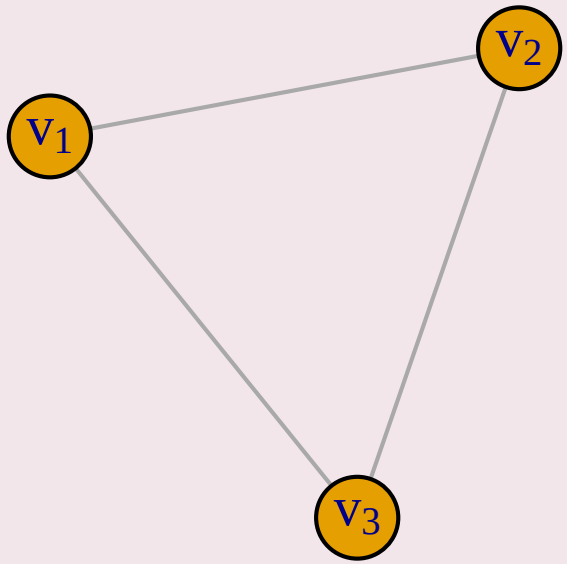


Figure: Stationary distribution example.

Stationary Distribution Solution

We are able to find the stationary distribution, when given an undirected graph. Before we present the solution, we must first define the Volume of a graph.

Graph Volume

Volume of an Undirected Graph

The volume of an undirected graph $G = (V, E)$ is equal to the sum of the degree of all vertices.

$$\text{vol}(G) = \sum_{v \in V} \deg(v)$$

Volume of a Directed Graph

The volume of a directed graph $G = (V, E)$ is equal to the sum of the in or out degree of all vertices.

$$\text{vol}(G) = \sum_{v \in V} \text{indeg}(v) = \sum_{v \in V} \text{outdeg}(v)$$

Note that the sum of all vertices in degree is equal to the out degree sum since each edge in the graph has one in and one out end.

Undirected Graph Stationary Distribution

Let $\vec{p}$ be a vector containing the probability of arriving at each vertex after a random walk, where

$$\vec{p} = \left[\frac{\deg(v_1)}{\text{vol}(G)} \frac{\deg(v_2)}{\text{vol}(G)} \cdots \frac{\deg(v_{\|V\|})}{\text{vol}(G)} \right]$$

Using the transition probability matrix T, we can take a random step on the graph by multiplying $T\vec{p}$. If we examine the probability of moving to vertex v_j:

$$\begin{aligned}
(T\vec{p})_i &= \sum_{v_j \to v_i} p_{v_i} T_{v_i, v_j} \\
&= \sum_{v_j \to v_i} \frac{\deg(v_j)}{\text{vol}(G)} \frac{1}{\deg(v_j)} \\
&= \sum_{v_j \to v_i} \frac{1}{\text{vol}(G)} = \frac{\deg(v_i)}{\text{vol}(G)} = p_i
\end{aligned}$$

Undirected Graph Stationary Distribution

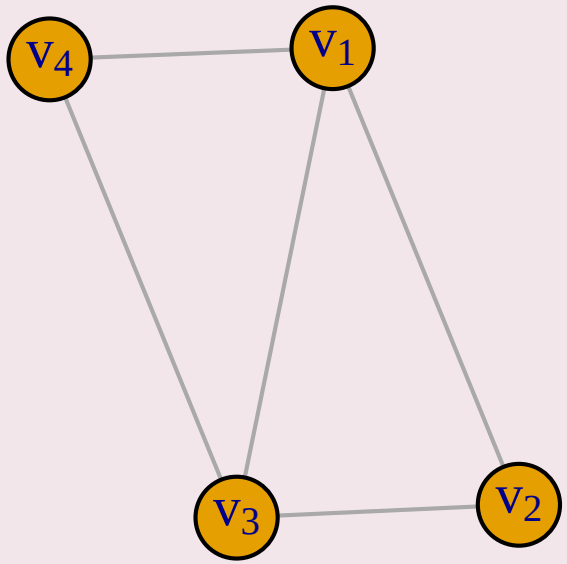
So, if we begin with the vertex state probability $\vec{p}$, given by

$$\vec{p} = \left[\frac{\deg(v_1)}{\text{vol}(G)} \; \frac{\deg(v_2)}{\text{vol}(G)} \cdots \frac{\deg(v_{\|V\|})}{\text{vol}(G)} \right]$$

and take a random step, the vertex state probability does not change. Therefore, the above defined vector $\vec{p}$ is the stationary distribution of undirected graph G.

Undirected Graph Stationary Distribution

Example

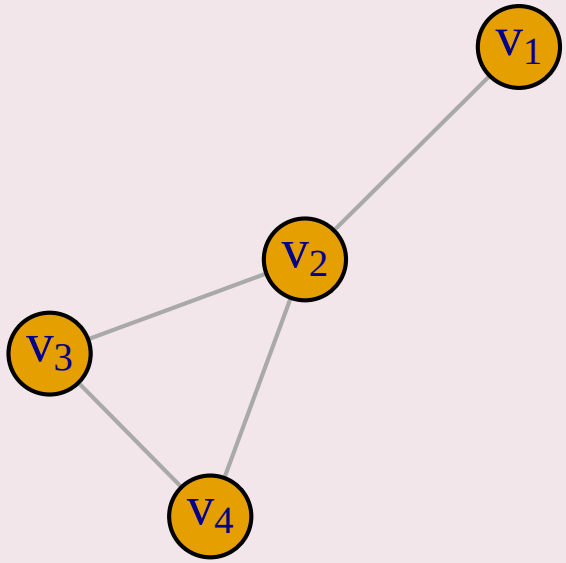


For the following graph:

1. Find the transition probability matrix T
2. Compute the stationary distribution $\vec{p}$
3. Show that $\vec{p} = T\vec{p}$

Undirected Graph Stationary Distribution

Problem



For the following graph:

1. Find the transition probability matrix T
2. Compute the stationary distribution $\vec{p}$
3. Show that $\vec{p} = T\vec{p}$

Outline

Random Walk on Directed Graphs

We have seen how to perform a random walk on an undirected graph and compute the stationary distribution.

We will now examine random walks on directed graphs.

The edges in a directed graph are one way, meaning that when performing the random walk, we may only follow edges in the direction they are pointing.

Example Random Walk on a Directed Graph

Two random walks from v_1 of length 3

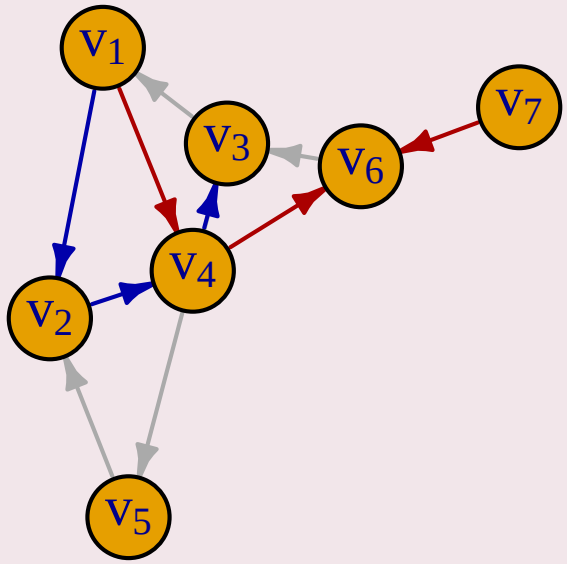


Figure: Random walk on a directed graph starting from v_1.

The blue random walk is legal, the red random walk is illegal since it does not obey the edge directions.

Probability Transition Matrices for Directed Graphs

The probability transition matrix for a random walk on a directed graph is constructed just as when using an undirected graph.

When adding the probabilities to the matrix, we must take into account the direction of the edges.

Probability Transition Matrices for Directed Graphs

Example

Construct the transition probability matrix for the following graph:

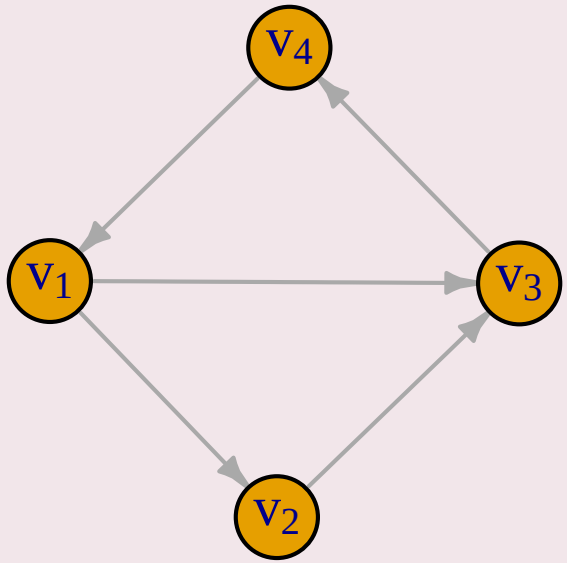


Figure: Directed probability transition matrix example.

Probability Transition Matrices for Directed Graphs

Problem

Construct the transition probability matrix for the following graph:

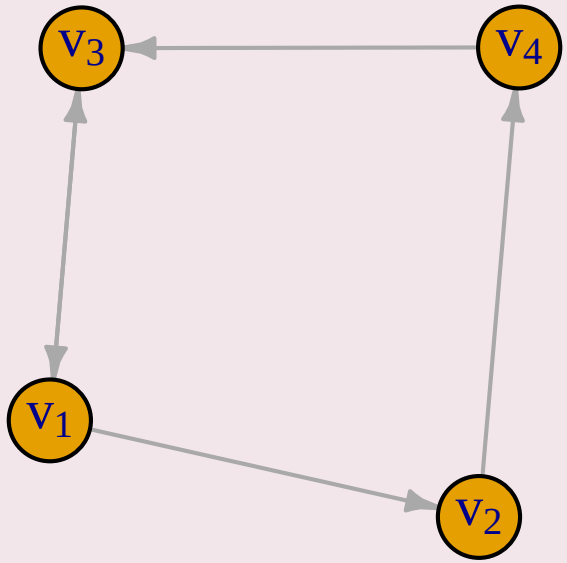


Figure: Directed probability transition matrix problem.

Stationary Distribution for Directed Graphs

We saw that we can easily find the stationary distribution for undirected graphs. Unfortunately there is no closed form solution for directed graphs.

We know the stationary distribution $\vec{p}$ satisfies the equation:

$$\vec{p} = T\vec{p}$$

where $\vec{p}$ is the stationary distribution and T is the probability transition matrix.

We can find $\vec{p}$ using either:

- Eigenvalue decomposition
- Power method

Eigenvalue solution

We have seen the eigenvalues being used in previous lectures, and now we have found another problem that can be solved using the eigenvalue decomposition!

Eigenvalue solution

We have seen the eigenvalues being used in previous lectures, and now we have found another problem that can be solved using the eigenvalue decomposition!

Given a square matrix A, the eigenvalue decomposition finds the eigenvectors $\vec{x}$ and the eigenvalues λ that satisfy:

$$\lambda\vec{x} = A\vec{x}$$

Eigenvalue solution

We have seen the eigenvalues being used in previous lectures, and now we have found another problem that can be solved using the eigenvalue decomposition!

Given a square matrix A, the eigenvalue decomposition finds the eigenvectors $\vec{x}$ and the eigenvalues λ that satisfy:

$$\lambda\vec{x} = A\vec{x}$$

If we find the eigenvalue decomposition of T, then the eigenvector with eigenvalue $\lambda = 1$ is the stationary distribution.

Using R, we can compute the eigenvalue decomposition of matrix `T` and store the result in `p` using:

```
p = eigen(T)
```

The eigenvalues are stored in `p$values` and the vectors in `p$vectors`.

Power Method

An alternative to using the eigenvalue decomposition for finding the stationary distribution, is to use the Power Method.

We know that the stationary distribution $\vec{p}$ is the state distribution of a random walk after an infinite number of steps. Luckily, the state distribution converges to the stationary distribution after a finite number of steps. Therefore, to find the stationary distribution, we perform a random walk, and stop the walk when the state distribution stops changing (has converged).

The power method algorithm is:

```
while(p != p.new) {
    p = p.new
    p.new = T %*% p
}
```

The stationary distribution is found in `p`.

Examining the Stationary Distribution

Example

Let's use R to compute the stationary distribution for the following graph:

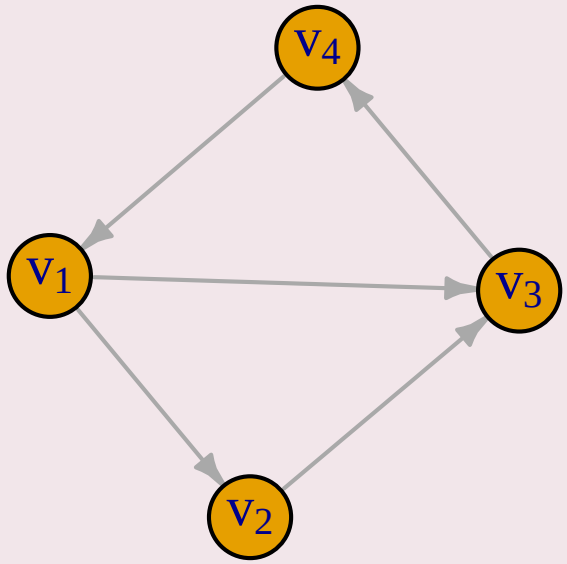


Figure: Stationary distribution example.

Which vertex do you think will have the greatest probability in the stationary distribution?

Outline

PageRank vs Graph Stationary Distribution

PageRank is a score given to a Web page based on its indegree. It is very similar to the stationary distribution over a directed graph.

The stationary distribution does not exist for certain directed graphs. The method used to compute PageRank, alters the graph to ensure that there is always a stationary distribution.

Ergodicity

Definition

Ergodic: Relating to or denoting systems or processes with the property that, given sufficient time, they include or impinge on all points in a given space and can be represented statistically by a reasonably large selection of points

A graph is *ergodic* if we can find a finite path between all pairs of vertices. The right graph is not ergodic (there is no path to v_1).

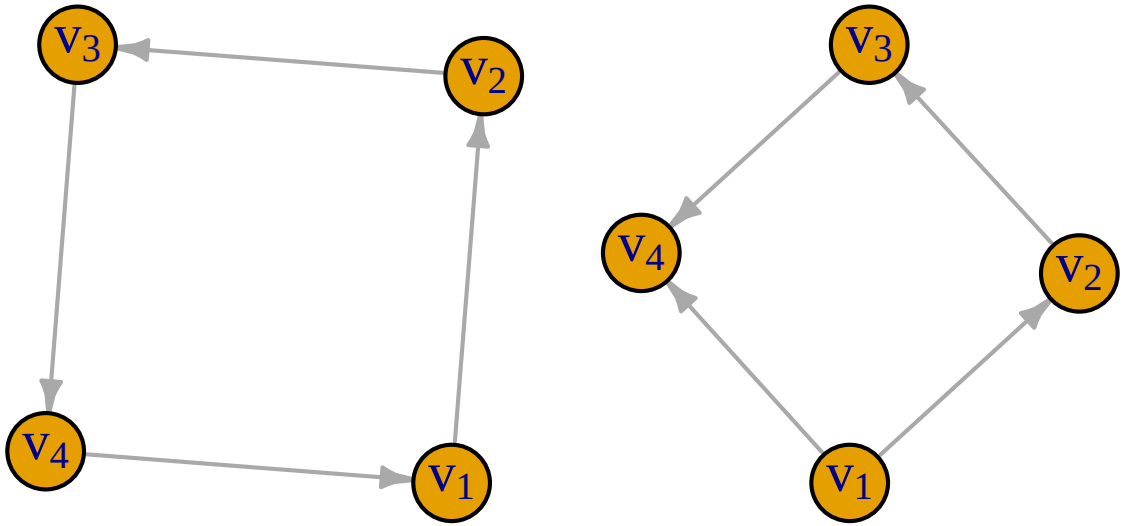


Figure: One ergodic, and one non-ergodic graph.

Ergodic Graphs

Problem

Which of the following graphs are ergodic?

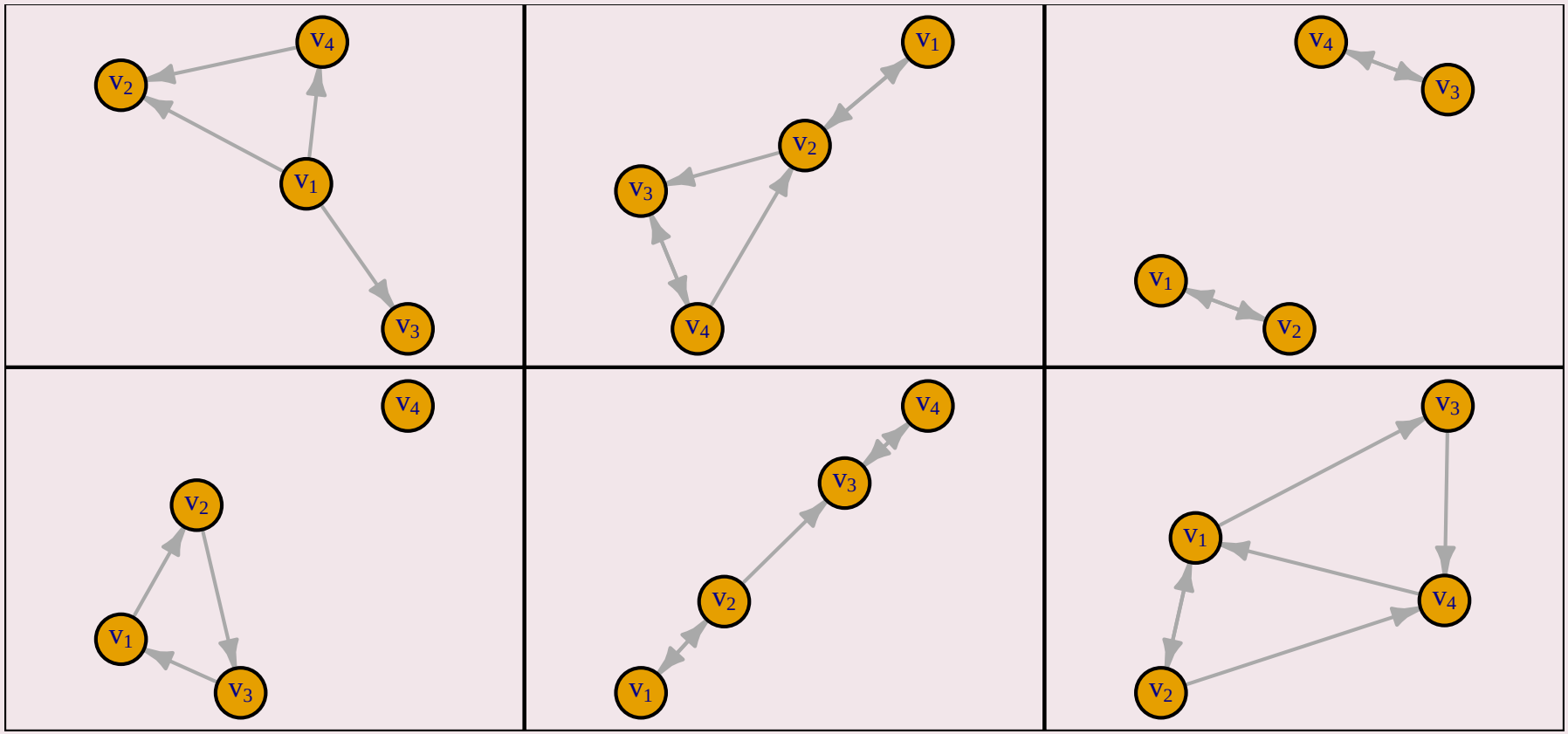


Figure: Find the ergodic graphs.

Non-Ergodicity

If a graph is non-ergodic, it is usually because there are vertices with only outgoing or incoming edges. These vertices are a problem when computing the stationary distribution, since:

- once a random walk leaves a vertex with only outgoing edges, it can never come back,
- once a random walk enters a vertex with only incoming edges, it can never leave.

So vertices with only outgoing edges obtain 0 probability in the stationary distribution, and vertices with only incoming edges force all other vertices to have 0 probability in the stationary distribution.

Non-Ergodicity

If a graph is non-ergodic, it is usually because there are vertices with only outgoing or incoming edges. These vertices are a problem when computing the stationary distribution, since:

- once a random walk leaves a vertex with only outgoing edges, it can never come back,
- once a random walk enters a vertex with only incoming edges, it can never leave.

So vertices with only outgoing edges obtain 0 probability in the stationary distribution, and vertices with only incoming edges force all other vertices to have 0 probability in the stationary distribution.

Remember from the language model lecture that we do not like zero probability! How do we remove zero probabilities? We smooth the distribution (just as we did for the language model distributions).

Smoothed Probability Transition Matrix

Recall that to smooth a distribution p, we require a background distribution b (what we expect the probabilities to be) and smoothing constant $\lambda \in [0, 1]$.

The smoothed distribution is:

$$s = \lambda p + (1 - \lambda) b$$

- If $\lambda = 1$, then $s = p$
- If $\lambda = 0$, then $s = b$
- If $0 < \lambda < 1$, then s is between p and b

So we want λ to be close to 1, but not 1.

Smoothed Probability Transition Matrix

Recall that to smooth a distribution p, we require a background distribution b (what we expect the probabilities to be) and smoothing constant $\lambda \in [0, 1]$.

The smoothed distribution is:

$$s = \lambda p + (1 - \lambda) b$$

- If $\lambda = 1$, then $s = p$
- If $\lambda = 0$, then $s = b$
- If $0 < \lambda < 1$, then s is between p and b

So we want λ to be close to 1, but not 1.

We have the probability of each transition T, but what do we choose as the background probability B and the smoothing parameter λ?

Random Surfer Model

The random surfer model was developed by Google to compute the stationary distribution of the set of Web pages on the Web graph.

To avoid the problems caused by non-ergodic graphs, the probability of moving from Web page to Web page (vertex to vertex) is given as:

- Follow link: the random surfer follows a Web page link with probability λ
- Random jump: the random surfer jumps to a random Web page with probability $1 - \lambda$

The random jump ensures that there is a probability of entering and leaving all pages, making the graph ergodic.

Random Surfer Probability Transition Matrix

For a given graph G with probability transition matrix T, the random surfer probability transition matrix is:

$$\lambda T + (1 - \lambda)B$$

where the background probability matrix B is:

$$B = \begin{bmatrix} 1/N & 1/N & \dots & 1/N \\ 1/N & 1/N & \dots & 1/N \\ \vdots & \vdots & \ddots & \vdots \\ 1/N & 1/N & \dots & 1/N \end{bmatrix}$$

and $N = ||V||$.

Note that the background distribution is Uniform (all pages have the same probability of $1/N$).

PageRank

PageRank is used (maybe) by Google to weight search results in terms of popularity.

The PageRank scores are the probabilities in the stationary distribution of the Random Surfer probability transition matrix from the set of Web pages.

A Web page's PageRank is found in the pages position in the stationary distribution.

Smoothed Probability Transition Matrix

Example

1. Compute the random surfer probability transition matrix for the following graph, using $\lambda = 0.8$:
2. Use R to compute the PageRank (stationary distribution) for the set of vertices.

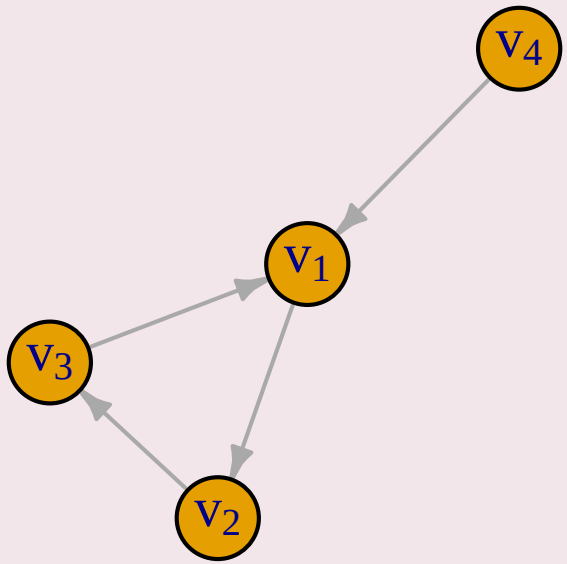


Figure: Smoothed probability transition matrix example.

Smoothed Probability Transition Matrix

Problem

1. Compute the random surfer probability transition matrix for the following graph, using $\lambda = 0.8$:
2. Verify that the stationary distribution is $\vec{p} = [\,0.47\ 0.05\ 0.05\ 0.43\,]$

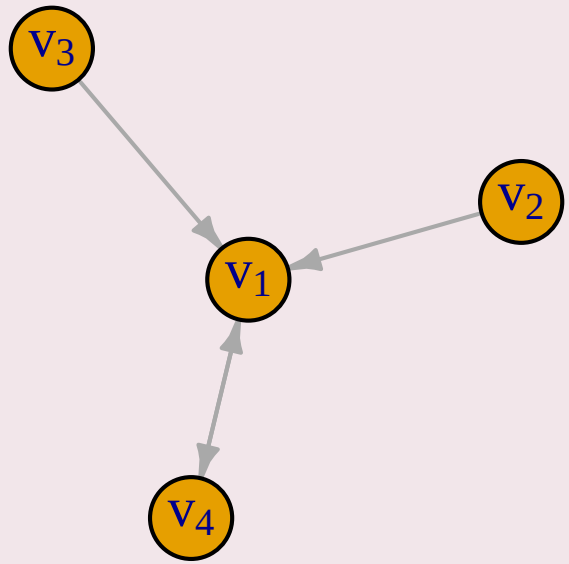


Figure: Smoothed probability transition matrix problem.

Summary

In this lecture, we found:

- A random walk on a graph moves from vertex to vertex, where each pair of vertices is connected with an edge.
- A state distribution provides us with the probability of landing on each vertex after a random walk of length n.
- A probability transition matrix provides us with the state distribution after a single step in a random walk.
- The stationary distribution provides us with the state distribution after an infinite length random walk. This distribution is similar to a measure of centrality.
- We can compute the stationary distribution for undirected and directed graphs.
- PageRank is the stationary distribution from the Random Surfer probability transition method.

Next Lecture

Time 1 (trends, trend periodicity)