Math 22 – Linear Algebra and its applications

- Lecture 28 -

Instructor: Bjoern Muetzel

GENERAL INFORMATION

- **Office hours:** Tu 1-3 pm, **Th**, Sun 2-4 **pm** in **KH 229**

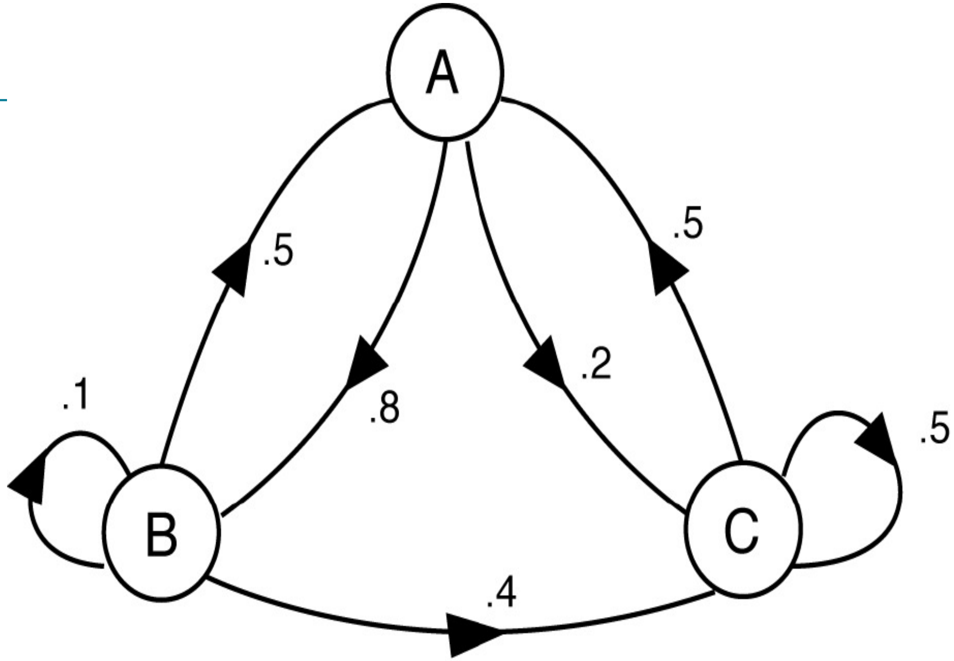
  Tutorial: Tu, **Th**, Sun **7-9 pm** in **KH 105**

- **Homework 8**: due **Wednesday** at **4 pm** outside **KH 008.** Please give in **part B, C** and **D.** There is **no part A.**

- **x-hour:** If we can not finish **Lecture 27** today, there will be a lecture during the x-hour, tomorrow, **Thursday.**

Applications

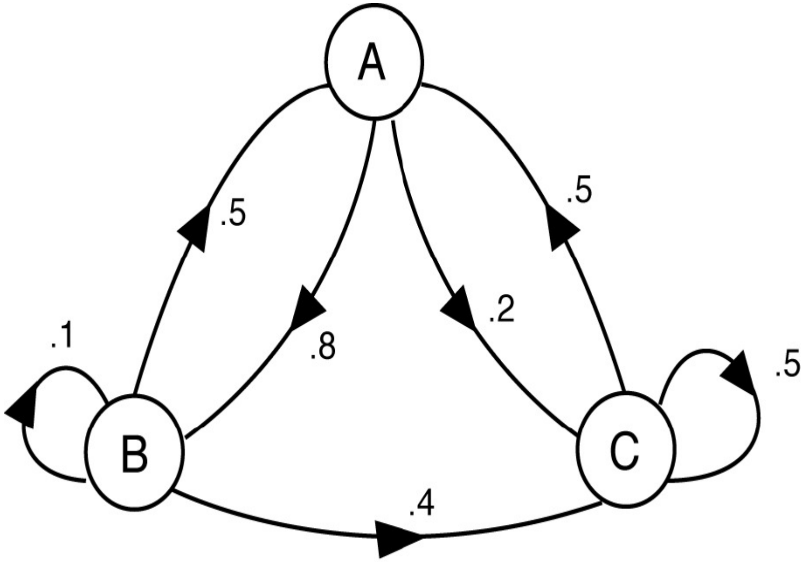
MARKOV CHAINS AND GOOGLE’S PAGE RANK ALGORITHM



Markov graph of transiton probabilites between states A, B and C

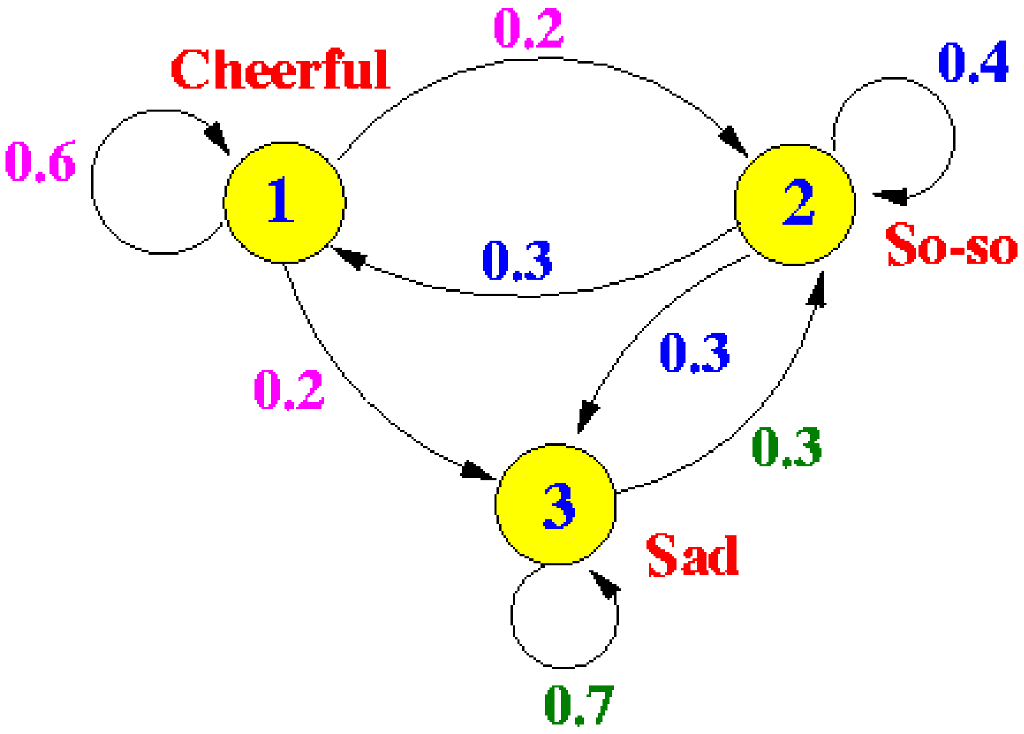
Summary:

Transitions or **flows** in **networks** can be analyzed by writing the information into a **matrix**. Finding the **steady state** of the system amounts to finding an **eigenvector of this matrix**.



Markov graph of transiton probabilites between states A, B and C

EXAMPLE: IN THE MOOD



MARKOV CHAINS

- **Definition: 1.)** A **probability vector** $\boldsymbol{v}$ in $\mathbb{R}^n$ is a vector with non-negative entries (probabilities) that add up to 1.

$$v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}, \boxed{v_1 + v_2 + \cdots + v_n = 1}, \quad \text{especially} \quad v_i \text{ in } [0,1].$$

2.) A **stochastic matrix** $\boldsymbol{P}$ is an n × n matrix whose columns are probability vectors.

3.) A **Markov chain** is a **sequence of probability vectors** $\boldsymbol{(x_k)_{k \text{ in } \mathbb{N}}}$, together with a stochastic matrix P, such that $\boldsymbol{x_0}$ **is the initial state** and

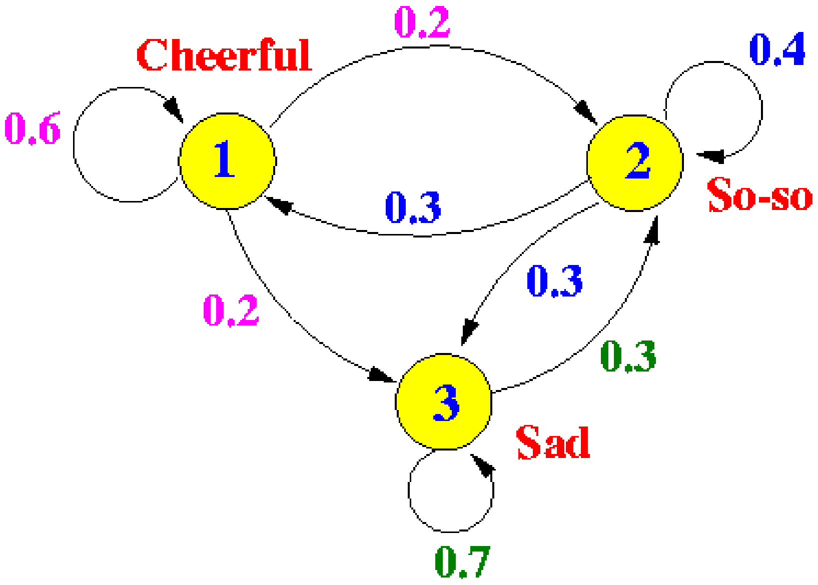
$$\boxed{\boldsymbol{x_k = P^k x_0}} \text{ or equivalently } \boxed{\boldsymbol{x_k = P x_{k-1}}} \quad \text{for all } \boldsymbol{k} \ \textbf{in } \mathbb{N}\backslash\{\boldsymbol{0}\}.$$

4.) A vector $\boldsymbol{x_k}$ of a Markov chain is called a **state vector.**

Visualization: We can visualize Markov chains with directed graphs.

- **Vertices** of the graph represent the **states** (entries of a vector).
- **Arrows** of the graph represent the **transitions** and their probability.
- We can write the transition probabilities into a stochastic matrix *P*.

Example:



$$\begin{array}{c|ccc|c}
\underline{\textbf{From:}} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \underline{\textbf{To:}} \\
 & 0.6 & 0.3 & 0 & \mathbf{1} \\
P= & 0.2 & 0.4 & 0.3 & \mathbf{2} \\
 & 0.2 & 0.3 & 0.7 & \mathbf{3}
\end{array}$$

MARKOV CHAINS

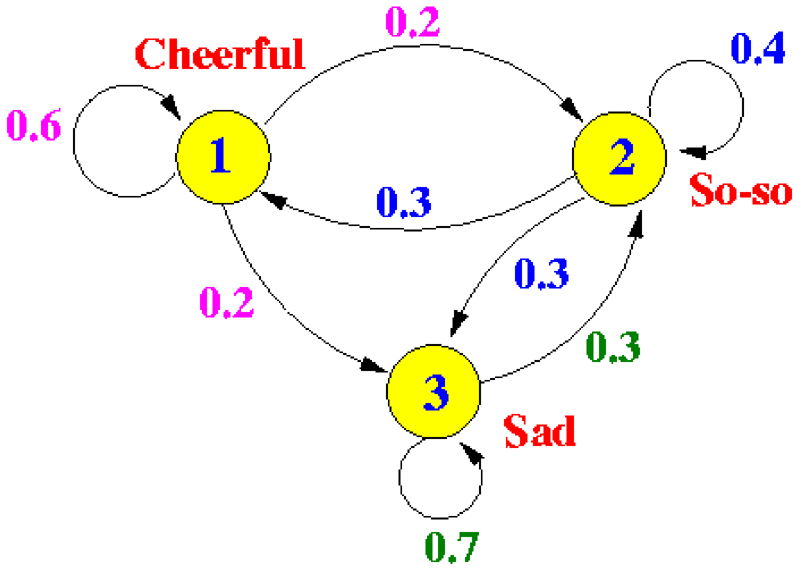
Visualization:

- The index **k** of a state vector $\boldsymbol{x_k}$ represents the **time interval**. $\boldsymbol{x_k}$ describes the state of the system at the time interval **k**.
- If $\boldsymbol{x_k = (x_{1k}, x_{2k}, \ldots, x_{nk})^T}$, then the entry $\boldsymbol{x_{ik}}$ describes the **probability of being in state i** at time **k**.
- The **transition probabilities** are **fixed** and show how the system progresses in the next time interval. This is equal to $\boldsymbol{x_{k+1} = Px_k}$.

Example:

$$P = \begin{bmatrix} 0.6 & 0.3 & 0 \\ 0.2 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.7 \end{bmatrix}, x_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$x_1 = \begin{bmatrix} 0.6 & 0.3 & 0 \\ 0.2 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.7 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.6 \\ 0.2 \\ 0.2 \end{bmatrix}$$



- **Definition:** Let *P* be an n × n stochastic matrix. A **steady-state** or **equilibrium vector q in** $\mathbb{R}^n$ is a vector, such that

$$\boxed{\boldsymbol{Pq} = \mathbf{1} \cdot \boldsymbol{q} = \boldsymbol{q}}.$$

Note: $\boldsymbol{q}$ is an **eigenvector** of *P* for the **eigenvalue 1**. Every stochastic matrix has a steady state vector.

Proof:

MARKOV CHAINS

- **Definition:** Let P be an n × n stochastic matrix. Then P is **regular** if some matrix power P^k contains **no zero entries**.

Theorem 1: (Markov chains) If P be an n × n **regular** stochastic matrix**,** then P has a **unique** steady-state vector $\boldsymbol{q}$ that is a probability vector. Furthermore, if $\boldsymbol{x_0}$ is **any initial state** and

$$\boxed{\boldsymbol{x_k = P^k x_0}} \text{ or equivalently } \boxed{\boldsymbol{x_k = P x_{k-1}}},$$

then the Markov chain

$(\boldsymbol{x_k})_{\boldsymbol{k} \textbf{ in } \mathbb{N}}$ converges to $\boldsymbol{q}$ as $k \to \infty$.

SEARCH ENGINES

How can we search the web for information?

This part of the presentation is based on the article *"How Google finds your needle in the haystack"* by David Austin.

Goal: Describe Google's PageRank Algorithm.

The web contains more than 30 trillion web pages.

How can we search these pages for information within seconds?

SEARCH ENGINES

What does a search engine do?

1.) **Index web pages:** Search the web and locate all web pages with public access and index the data on these pages.

2.) **Rank the importance of pages:** In order to display the most relevant pages first, it needs to decide which page is most important.

3.)**Match search criteria:** When a user enters one or several keywords, the search engine matches it to the indexed pages with the same keywords. Among these it picks the **most important** ones and displays them.

SEARCH ENGINES

What does a search engine do?

Google has indexed more than **30 trillion pages**. Most of these contain about 10.000 words. This means that there is a huge number of pages that contain the words of a search phrase.

The Big Problem:

Rank the pages such that the important ones are displayed first.

Idea: Model webpages with links as a directed graph. Calculate the importance of a page according to the number of pages linking to it.

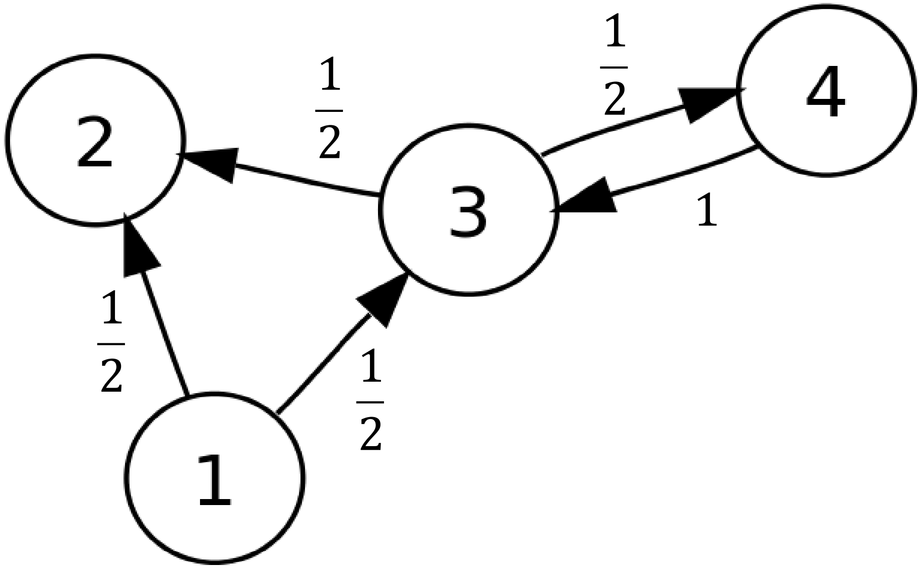
SEARCH ENGINES

Model: We model the web as a directed graph:

- Every vertex $(w_i)_i$ is a webpage.
- Every link from one page to another is an arrow.
- Each webpage distributes its whole value onto the pages its links to with equal weight. Self-links are not counted.

Example:

$$\begin{array}{cccccc} \underline{\textbf{From:}} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \underline{\textbf{To:}} \\ & 0 & 0 & 0 & 0 & \mathbf{1} \\ \text{H=} & 0.5 & 0 & 0.5 & 0 & \mathbf{2} \\ & 0.5 & 0 & 0 & 1 & \mathbf{3} \\ & 0 & 0 & 0.5 & 0 & \mathbf{4} \end{array}$$



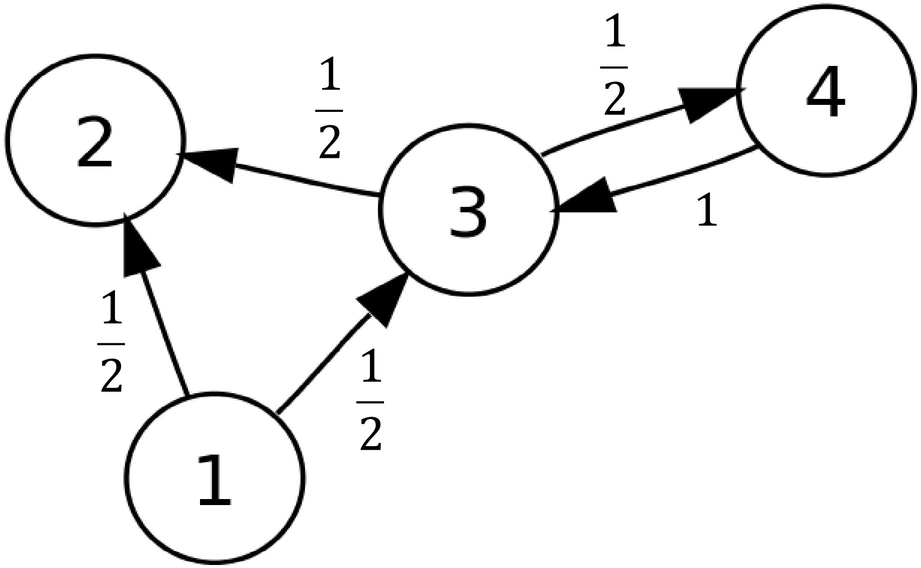
HYPERLINK MATRIX

Definition: Let $(w_i)_i$ be the set of webpages on the internet. Let $\boldsymbol{n_i}$ be the **number of pages** the page w_i links to. The **hyperlink matrix** $\boldsymbol{H} = (\boldsymbol{h_{ij}})_{\boldsymbol{i,j}}$ is defined by

$$h_{ij} = \begin{cases} 1/n_j & \text{if} \quad w_j \text{ links to } w_i \\ 0 & \text{otherwise .} \end{cases}$$

Example:

$$\begin{array}{rccccl} \underline{\textbf{From:}} & \textbf{1} & \textbf{2} & \textbf{3} & \textbf{4} & \underline{\textbf{To:}} \\ & 0 & 0 & 0 & 0 & \textbf{1} \\ \text{H=} & 0.5 & 0 & 0.5 & 0 & \textbf{2} \\ & 0.5 & 0 & 0 & 1 & \textbf{3} \\ & 0 & 0 & 0.5 & 0 & \textbf{4} \end{array}$$



HYPERLINK MATRIX

If H was a regular stochastic matrix we could use Theorem 1:

Theorem 1: (Markov chains) If P be an $n \times n$ regular stochastic matrix, then P has a **unique** steady-state vector $\boldsymbol{q}$ that is a probability vector. If $\boldsymbol{x_0}$ is **any initial state** and $\boldsymbol{x_k = P^k x_0}$ then the Markov chain $(\boldsymbol{x_k})_{\boldsymbol{k} \textbf{ in } \mathbb{N}}$ converges to $\boldsymbol{q}$ as $k \to \infty$.

This means that we could assign the **importance** $\boldsymbol{q_i}$ of a webpage $\boldsymbol{w_i}$ via the steady state vector $\boldsymbol{q} = (\boldsymbol{q_1}, \boldsymbol{q_2}, \dots, \boldsymbol{q_k}, \dots)$, such that

$$\boldsymbol{Hq = q}.$$

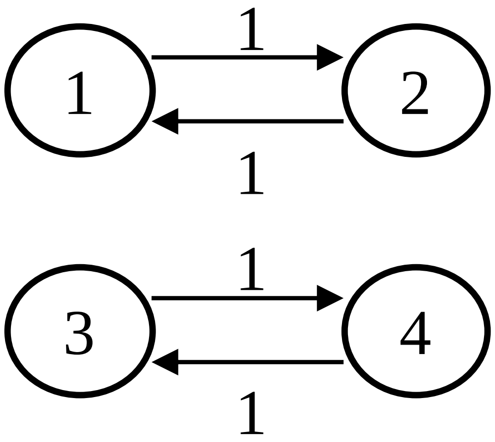
This means that the importance would be the steady-state result of a repeated process of assigning importance.

How can we turn H into a regular stochastic matrix?

Problem 1: Web pages without links (see previous example). In this case there is a **zero column** in the matrix H.

Problem 2: Groups of pages that do not link to other groups.

$$H=\left[\begin{array}{cc|cc} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{array}\right]$$



In this case the matrix can **not** be **regular**, as no power of H can have nonzero entries. There would be **no unique steady state vector**.

GOOGLE MATRIX

How can we turn H into a regular stochastic matrix?

Solution: Let **S** be the matrix where zero columns in H are replaced by a column **(1/n,1/n,…,1/n),** where **n** is the number of webpages.

Definition: We define the **Google matrix G** to be

$$\boxed{\mathrm{G} = 0.85 \cdot \boldsymbol{S} + 0.15 \cdot \frac{1}{n}\mathbf{1}} \;, \qquad \text{where}$$

$\boxed{\mathbf{1}}$ is the matrix with all entries equal to 1.

Note: With this definition G is a regular stochastic matrix. We can apply **Theorem 1** and define an importance for each webpage.

How can we define the importance of a webpage ?

The Google matrix G is a regular stochastic matrix. We define

Definition: The **importance** $\boldsymbol{q_i}$ of a webpage $\boldsymbol{w_i}$ is the entry in the steady state vector $\boldsymbol{q} = (\boldsymbol{q_1}, \boldsymbol{q_2}, \dots, \boldsymbol{q_k}, \dots)$, such that

$$\boxed{\boldsymbol{Gq = q}} \; .$$

We call **q** the **importance vector** of the Google matrix.

Note 1: The steady state vector $\boldsymbol{q}$ can be scaled to obtain a more practical range.

Note 2: In practice the calculation of $\boldsymbol{q}$ is done once per month using the iteration from **Theorem 1**: $\boxed{\boldsymbol{x_k = Gx_{k-1}}}$.