7. Die Google-Matrix

(7.1) Beispiel. Die Google-Matrix. Das Internet besteht aus Internetseiten P_i, $i = 1, \ldots, N$, (wobei $N \approx 1.3 \cdot 10^{14}$), die untereinander durch sogenannte Hyperlinks verbunden sind. Wir schreiben $P_j \to P_i$, wenn ein Hyperlink von der Seite P_j zur Seite P_i führt. Die Suchmaschine GOOGLE bewertet jede einzelne Seite P_i mit einer Zahl r_i („Pagerank“). Je größer r_i, desto besser die Seite. Die Bewertung einer Seite ist umso besser, je mehr gute Seiten auf sie verweisen, und zwar verteilt jede Seite ihren Pagerank gleichmäßig auf diejenigen Seiten, auf die sie verweist:

$$r_i = \sum_{\substack{j \\ P_j \to P_i}} \frac{r_j}{d_j}$$

wobei d_j die Anzahl der von der Seite P_j ausgehenden Hyperlinks bezeichnet. Mithilfe der Hyperlinkmatrix

$$H_{ij} = \begin{cases} \frac{1}{d_j} & \text{wenn } P_j \to P_i \\ 0 & \text{sonst} \end{cases}$$

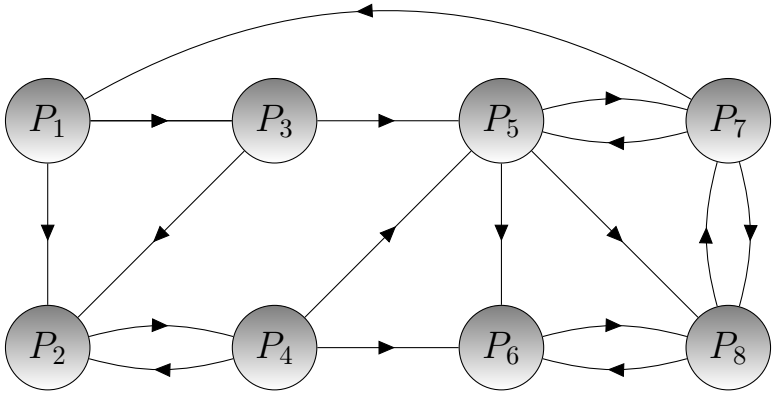
kann man diese Beziehung als Matrixprodukt

$$\vec{r} = H \cdot \vec{r}$$

schreiben, wobei der Spaltenvektor $\vec{r}$ die Pageranks enthält. Stellt man sich die Verteilung des Pageranks als Wasserströmung vor, dann entspricht die gesuchte Pagerankverteilung $\vec{r}$ einem dynamischen Gleichgewichtszustand. Es handelt sich um ein sogenanntes *Eigenwertproblem*.

Eine andere Interpretation ist die folgende: Angenommen, ein Leser startet auf einer zufällig gewählten Seite P_{i_0} und wählt dann zufällig einen der d_{i_0} Hyperlinks, landet auf einer Seite P_{i_1}, und so weiter. Dann ist der Pagerank r_i proportional zur Zeit, die auf der Seite P_i verbracht wird.

Zur Illustration ein kleines Beispiel:



Die dazugehörige Hyperlinkmatrix ist

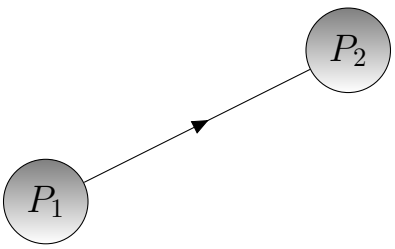
$$H = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 \\ 1/2 & 0 & 1/2 & 1/3 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 1/3 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/3 & 1 & 1/3 & 0 \end{pmatrix} \qquad \vec{r} = \begin{pmatrix} \frac{3}{50} \\ \frac{27}{400} \\ \frac{3}{100} \\ \frac{27}{400} \\ \frac{39}{400} \\ \frac{81}{400} \\ \frac{9}{50} \\ \frac{59}{200} \end{pmatrix} = \begin{pmatrix} 0.06 \\ 0.0675 \\ 0.03 \\ 0.0675 \\ 0.0975 \\ 0.2025 \\ 0.18 \\ 0.295 \end{pmatrix}$$

Lösungsansatz: Iteration $n \to \infty$:

$$\vec{r}(n+1) = H \cdot \vec{r}(n)$$

wobei $\vec{r}(0)$ ein beliebiger Anfangsvektor ist mit $r_i(0) \geq 0$ und $\sum_i r_i(0) = 1$.

Problem Nr 1: Seiten ohne ausgehende Hyperlinks schlucken „zuviel Wahrscheinlichkeit“, z.B.



Lösung: Wenn man in einer Sackgasse landet, wählt man die nächste Seite *zufällig*, d.h., irgendeine Seite P_i wird mit Wahrscheinlichkeit $1/N$ gewählt.

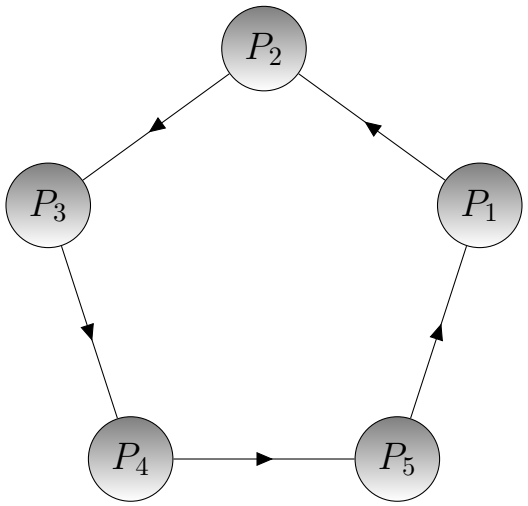
Wir müssen also unsere Matrix modifizieren

$$S = H + \tfrac{1}{N}A$$

wobei

$$a_{ij} = \begin{cases} 1 & \text{wenn } P_j \text{ eine Sackgasse ist} \\ 0 & \text{sonst} \end{cases}$$

Problem Nr 2: Es kann immer noch passieren, dass der Prozess nicht konvergiert, z.B.



Lösung: Eine weitere Modifikation: Der Leser kann sich jederzeit, also auch auf einer Seite mit Hyperlinks entscheiden, auf eine zufällig gewählte neue Seite zu springen:

$$G = \alpha S + (1-\alpha)\tfrac{1}{N}\mathbf{1}$$

wobei $\mathbf{1}$ die Matrix mit lauter 1 ist. Interpretation: mit Wahrscheinlichkeit α öffnet der Leser einen Hyperlink auf der gerade besuchten Seite und mit Wahrscheinlichkeit $1-\alpha$ springt er auf eine zufällig gewählte andere Seite.

Die Matrizen H, S und G sind *stochastisch*, d.h., sie haben nichtnegative Einträge und die Spaltensummen sind 1. Mithilfe der Theorie von *Perron-Frobenius* kann man zeigen, dass für $\alpha < 1$ die Iteration

$$\vec{r}(n+1) = G \cdot \vec{r}(n)$$

für jeden beliebigen Startvektor $\vec{r}(0)$ mit positiven Einträgen gegen eine Lösung der Gleichung

$$\vec{r} = G \cdot \vec{r}$$

konvergiert. Die Lösung ist bis auf einen skalaren Faktor eindeutig.

S. Brin und L. Page wählten $\alpha = 0.85$ und die Iteration stabilisiert sich nach 50-100 Iterationen. Diese Rechnung wird angeblich ca. ein Mal pro Monat ausgeführt.