Übung 10

Page Rank

Page Rank

- Wichtigkeit von Webseiten wird nach den Links bewertet, die auf die Webseite verweisen
- Je wichtiger die Webseite ist, von der der Link ausgeht, desto mehr zählen seine ausgehenden Links

Random Surfer Modell

- Es existiert ein Web Surfer, der zu jedem Zeitpunkt t auf einer Webseite ist
- Der Surfer folgt mit Wahrscheinlichkeit β einen zufällig ausgewählten Link auf der Webseite
- Mit Wahrscheinlichkeit $(1 - \beta)$ springt er auf eine zufällige Webseite

10-1a

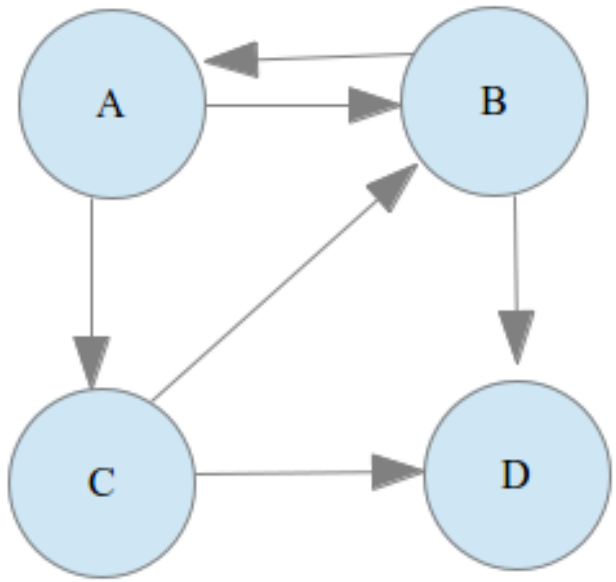
Wie vermeidet es der PageRank Algorithmus in einem "Dead End" hängenzubleiben?

Dead Ends

- Dead Ends sind Seiten die keine ausgehenden Links haben
- Der Algorithmus verhindert das Hängenbleiben durch sogenannte Teleports (d. h. es wird auf eine zufällige Seite gesurft)
- Dabei ist die Wahrscheinlichkeit, dass auf eine bestimmte Seite gesurft wird $\frac{1}{n}$ wobei n die Anzahl der erfassten Seiten ist

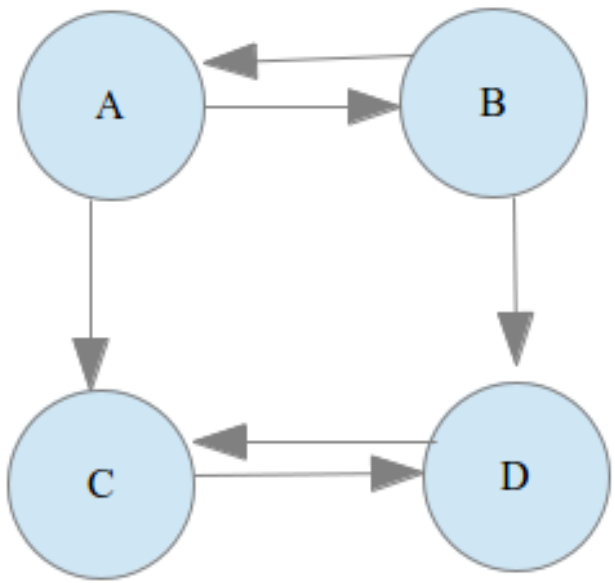
Dead Ends

- Knoten D ist ein Dead End
- Wenn der Algorithmus in D landet, besucht er als nächstes A, B, C oder D jeweils mit Wahrscheinlichkeit $\frac{1}{4}$



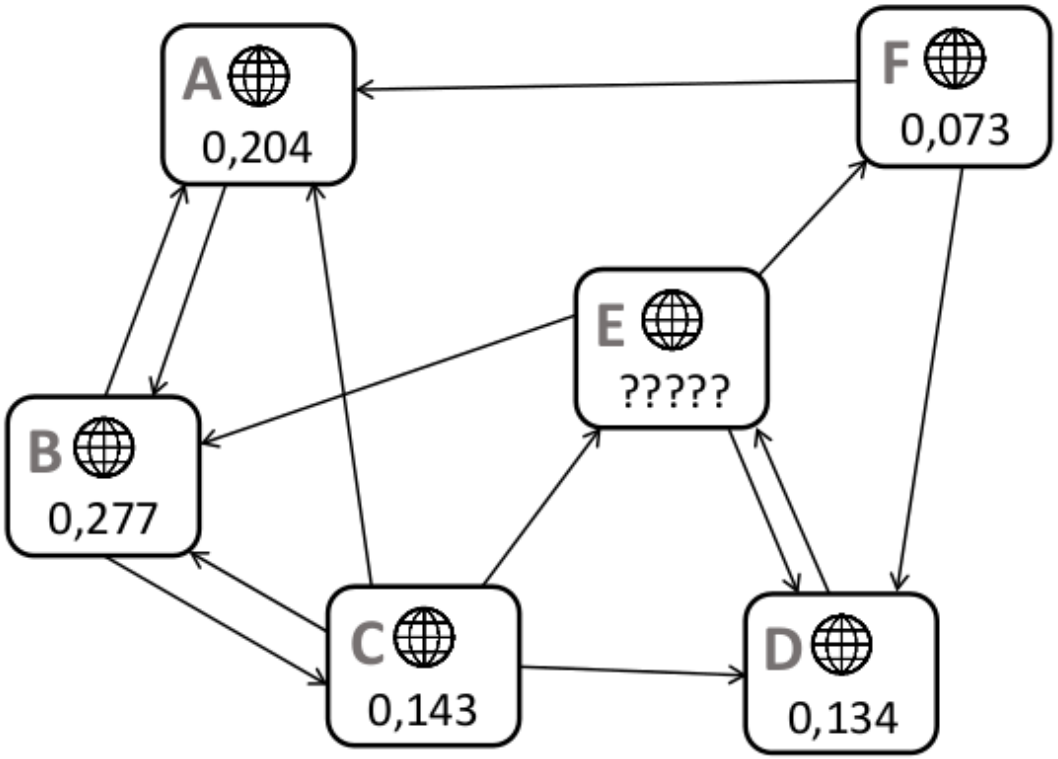
Spider Traps

- Alle ausgehenden Links von einer Gruppe von Seiten sind innerhalb der Gruppe
- C und D sind eine Spider Trap
- Um das zu verhindern, wird zu jedem Zeitpunkt einem Link nur mit Wahrscheinlichkeit β gefolgt
- Mit Wahrscheinlichkeit $(1 - \beta)$ wird zu zufälliger Seite gesprungen



10-1b

Berechne die Google Matrix zu dem Graphen mit $\beta = 0.85$



10-1b

- Bilde die Adjazenzmatrix und ihre Transponierte

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \quad A^T = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

10-1b

- Bilde Matrix M durch Normierung der Spalten von A^T

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

Diese Matrix ist bereits reell, nicht-negativ und spaltenstochastisch. Dies wäre im Falle eines Dead Ends nicht so, da wir dann eine 0-Spalte hätten. Dann müsste man in dieser Spalte alle Werte auf $\frac{1}{6}$ setzen.

10-1b

- Google Matrix $G = \beta \cdot M + \frac{1-\beta}{n} \cdot 1$

$$0.85 \cdot \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{4} & 0 & 0 & \frac{1}{2} \\ 1 & 0 & \frac{1}{4} & 0 & \frac{1}{3} & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 & \frac{1}{3} & \frac{1}{2} \\ 0 & 0 & \frac{1}{4} & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \end{bmatrix} + (1-\beta) \begin{bmatrix} \frac{1}{6} & \cdots & \frac{1}{6} \\ \vdots & \ddots & \vdots \\ \frac{1}{6} & \cdots & \frac{1}{6} \end{bmatrix} =$$

10-1b

- Google Matrix $G = \beta \cdot M + \frac{1-\beta}{n} \cdot 1$

$$= \frac{1}{60} \begin{bmatrix} 1.5 & 27 & 14.25 & 1.5 & 1.5 & 27 \\ 52.5 & 1.5 & 14.25 & 1.5 & 18.5 & 1.5 \\ 1.5 & 27 & 1.5 & 1.5 & 1.5 & 1.5 \\ 1.5 & 1.5 & 14.25 & 1.5 & 18.5 & 27 \\ 1.5 & 1.5 & 14.25 & 52.5 & 1.5 & 1.5 \\ 1.5 & 1.5 & 1.5 & 1.5 & 18.5 & 1.5 \end{bmatrix}$$

10-1c

Wie können die PageRank-Werte mit Hilfe der Google Matrix berechnet werden?

10-1c

- Google-Matrix ist stochastisch $\Rightarrow$ es existiert ein Eigenvektor von G zum Eigenwert 1
- $\Rightarrow$ bei dem Eigenwertproblem $G \cdot x = x$ ist der Vektor x ein stochastischer Vektor, der aus den PageRank Werten besteht
- Um den Eigenvektor x_i zum Eigenwert λ_i zu finden, kann man das Gleichungssystem $(G - \lambda_i E)x_i = 0$ lösen
- $\lambda_i = 1 \Rightarrow (G - E) = 0$

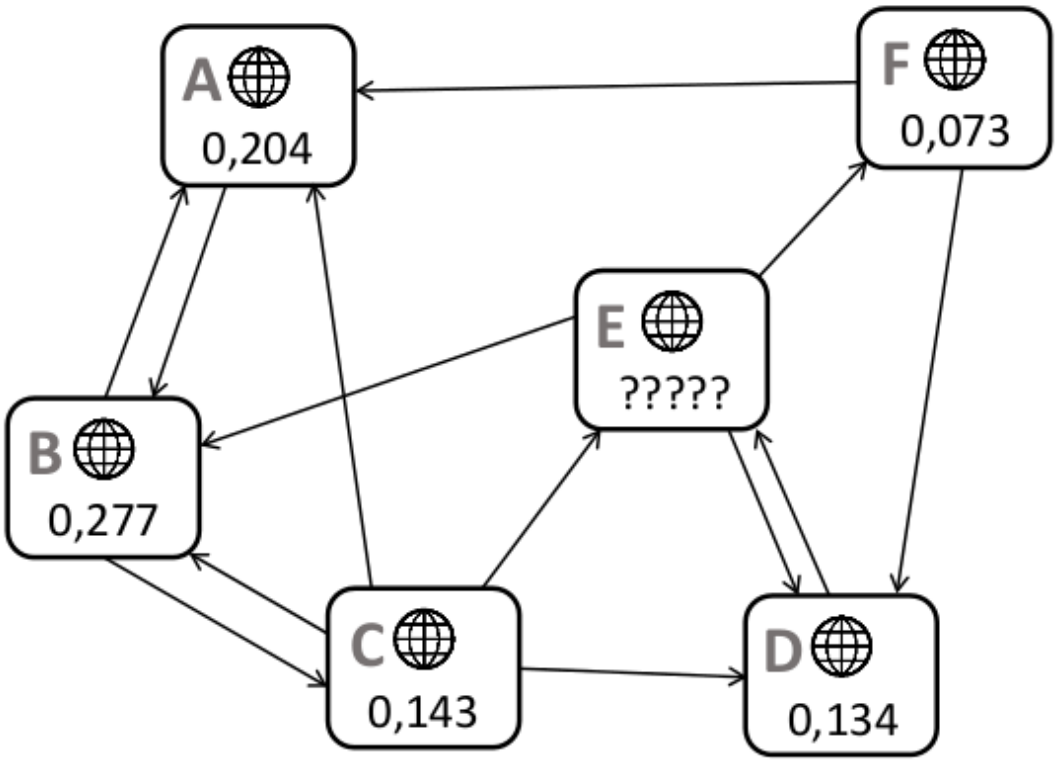
10-1c

Zu lösendes Gleichungssystem

$$\left[\begin{array}{cccccc|c} -0.975 & 0.450 & 0.238 & 0.025 & 0.025 & 0.450 & 0.000 \\ 0.000 & -0.571 & 0.451 & 0.047 & 0.331 & 0.429 & 0.000 \\ 0.000 & 0.000 & -0.605 & 0.064 & 0.293 & 0.383 & 0.000 \\ 0.000 & 0.000 & 0.000 & -0.943 & 0.462 & 0.662 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.376 & -0.873 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \end{array}\right]$$

10-1d

Berechnen Sie den fehlenden PageRank Wert des Knotens E im Graphen mit Hilfe der PageRank Gleichung



10-1d

$PR_E = \frac{1-\beta}{n} + \beta \sum_{i \to E} \frac{PR_i}{d_i}$
$PR_E = \frac{0.15}{6} + 0.85(\frac{0.143}{4} + 0.134)$
$PR_E = 0.169$
Berechnung wäre auch mit Hilfe des Gleichungssystems aus (c) möglich gewesen:
$0.376 \cdot PR_E + (-0.873) \cdot 0.073 = 0$
$PR_E = \frac{0.873-0.073}{0.376}$
$PR_E = 0.169$