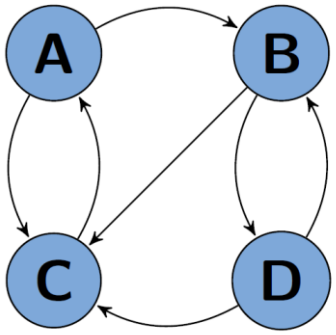
Übergangsprozesse-Prozesse
Google PageRank-Algorithmus

1 Larry Page und Andrej Brin haben an der Stanford-University den Page Rank-Algorithmus entwickelt und ihn patentieren lassen. 1998 gründeten sie das Unternehmen Google. Der Algorithmus bildet die Basis für die Bewertung von Web-Seiten, d.h. er entscheidet, in welcher Reihenfolge Google die Links zu einem Suchbegriff auflistet.

Die Grundidee besteht darin, sogenannte spiders oder webcrawler durch das Internet zu schicken. Es handelt sich dabei um Programme, die auf einer Seite die Basisdaten (Name der Seite, Tags etc. aufzeichnen, dann nach Links zu anderen Seiten suchen und per Zufallsauswahl zu einer der verlinkten Seiten springen.
Die Idee ist: Je wichtiger die Seite ist, desto mehr wichtige Seiten verlinken darauf. Eine Seite erhält also einen umso höheren Bewertungs-Rank, je häufiger dort spiders vorbeikommen.

Die Idee klingt vielleicht seltsam, war aber irgendwie was wert (Google bzw. der Mutterkonzern Alphabet ist mit 887,9Mrd Dollar das zweitteuerste Unternehmen der Welt – gemessen am Börsenwert). https://www.handelsblatt.com/finanzen/anlagestrategie/trends/apple-google-amazon-das-sind-die-zehn-wertvollsten-unternehmen-der-welt/22856326.html?ticket=ST-486936-UeWaTT5VsszERFl4rGjC-ap5

Einfaches Beispiel: Vier Internet-Seite A, B, C, D sind nach folgendem Muster miteinander verlinkt:



Da von A zwei Links weggehen, ergibt sich für den Übergang A→B eine Wahrscheinlichkeit von $\frac{1}{2}$ und für A→C eine von $\frac{1}{2}$.
Entsprechend geht man bei den anderen Seiten vor.

a) Stell die Übergangsmatrix M auf.

b) Bestimme M^4. Gib die Bedeutung des Elements rechts oben im Sachzusammenhang an.
Die Wahrscheinlichkeit, dass eine spider, die sich auf D befindet, 4 Runden später auf A gelandet ist, beträgt 0,375

100 spiders befinden sich auf Seite A – nirgendwo sonst ist eine spider.

c) Berechne die vermutliche Verteilung nach 3 (Übergangs-)Runden.

d) Bestimme die Wahrscheinlichkeit, dass eine bestimmte spider zuerst nach C wechselt, dann nach A, dann nach B, nach D, nach B, nach C und nach A.

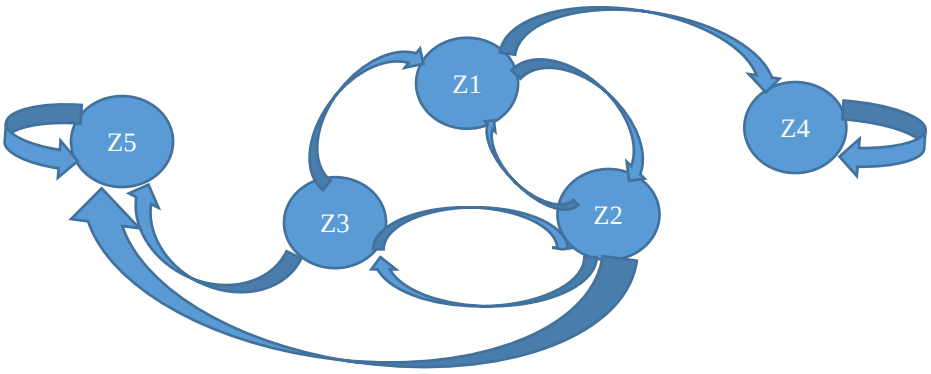
e) Berechne die Wahrscheinlichkeit, dass sich von den 100 spiders nach drei Runden zwischen 40 und 60 auf B befinden.

f) Untersuche M auf Eigenwerte und Eigenvektoren. Ermittle damit die Reihenfolge (also das PageRanking).

g) Du bist Webmaster der Seite A und möchtest im PageRank möglichst weit nach oben. Dazu darfst du genau einen Link im obigen Netz hinzufügen oder entfernen. Was tust du?

2 Natürlich läuft der PageRank-Algorithmus in Wirklichkeit nicht *genau* nach diesem Prinzip. Schon einfache Besonderheiten würden ihn dann vor Probleme stellen.

Gehe von folgendem Übergangsgraph aus:



a) Welche Zustände sind absorbierend?

b) Berechne die Wahrscheinlichkeit, dass eine bestimmte spider, die sich auf Z1 befindet, irgendwann bei Z4 endet.

c) Berechne die Wahrscheinlichkeit, dass von 15 spiders, die sich auf Z1 befinden, weniger als ein Drittel irgendwann bei Z5 enden.

d) Berechne die mittlere Wartezeit (Verweildauer), bis eine bestimmte spider, die bei Z1 startet auf einem absorbierenden Zustand gelandet ist.

e) Berechne die Wahrscheinlichkeit, dass eine bestimmte spider, die sich auf Z1 befindet, sich zwei Runden später wieder auf Z1 befindet **und** irgendwann bei Z5 endet.

f) Im Zusammenhang mit Webseiten taucht nicht nur das Problem auf, dass eine Website *auf sich selbst* verlinkt, sondern auch, dass sie *gar nicht* verlinkt. Schlage vor, wohin eine spider dann von dort aus gehen könnte oder welche Übergangswahrscheinlichkeiten dann gewählt werden könnten.