Suchmaschinen und Markov-Ketten

Zielstellung

1. Wir geben einen kurzen Überblick über die Arbeitsweise von **Suchmaschinen** für das Internet.
   - Eine Suchmaschine erwartet als Eingabe
     ein Stichwort oder eine Liste von Stichworten
   - und gibt als Ausgabe eine Liste von Links auf möglichst informative Webseiten zu diesen Stichworten. Die Liste soll so sortiert sein, dass

     die informativsten Links am „weitesten oben“ stehen.

2. Gleichzeitig geben wir eine Einführung in

   Markov-Ketten,

   einem wichtigen Werkzeug in der Modellierung einfacher stochastischer Prozesse.

Das Internet

Die von Suchmaschinen zu bewältigenden Datenmengen sind immens! Quelle (Zuletzt besucht am 09.12.2014.)

Danach gab es 2012

- 634 Millionen Websites,
  wobei 51 Millionen in 2012 hinzugekommen sind,
- 3,5 Milliarden Webseiten,
- 2,4 Milliarden Internetnutzer weltweit
- und eine 1,2 Billionen, also 10^{12} Suchanfragen auf Google allein.

Und dann müssen Suchanfragen auch noch in „Echtzeit“ beantwortet werden.

Die zentrale Aufgabe:

Bewerte den Informationsgehalt der Webseiten

in Bezug auf die jeweilige Kombination der Suchbegriffe!

Die Architekur von Suchmaschinen

Suchmaschinen: Die wichtigsten Komponenten

Anfragen für einen sich rasant ändernden Suchraum gigantischer Größe sind ohne merkliche Reaktionszeit zu beantworten.

(1) **Web-Crawler** durchforsten das Internet, um neue oder veränderte Webseiten zu identifizieren.

(2) Die von den Crawlern gefundenen Informationen werden in einer komplexen **Datenstruktur** gespeichert, um bei Eingabe von Suchbegriffen in „Echtzeit“ alle relevanten Webseiten ermitteln zu können.

(3) **Bewerte die Webseiten**

*hinsichtlich ihrer Relevanz für mögliche Suchbegriffe wie auch hinsichtlich ihrer **generellen** Bedeutung im Internet.*

Datenstrukturen für Suchmaschinen

Die Datenstruktur: Index und invertierter Index

1. Im **Index** werden alle vom Crawler gefundenen Webseiten w gespeichert:
   - URL (d.h. die Adresse) und Inhalt von w.
   - Der Inhalt von w wird analysiert: Alle vorkommenden Worte werden in Häufigkeit und Sichtbarkeit (Vorkommen in Überschriften, Schriftgöße etc.) erfasst.
   - Die auf w zeigenden Hyperlinks werden ebenfalls analysiert:
     - ⋆ Welche Begriffe tauchen in der Beschriftung des Links auf?
     - ⋆ Wie prominent ist der Link platziert?
2. Aus dem Index wird der **invertierte Index** erzeugt, der zu jedem möglichen Suchbegriff eine Liste aller Webseiten enthält, die den Suchbegriff enthalten.
   - Für jede in der Liste auftauchende Webseite w wird die Sichtbarkeit des Begriffs innerhalb von w und innerhalb der auf w zeigenden Seiten aufgeführt.
   - Mit diesen Zusatzinformationen und mit Hilfe ihrer

     „grundsätzlichen“ Bedeutung

     wird die Seite w in die Liste eingereiht.
     - ⋆ Wie die Einreihung erfolgt, ist Betriebsgeheimnis der Suchmaschinenbetreiber.

Wie bestimmt man **die grundsätzliche Bedeutung einer Webseite**?

Page-Rank mittels Peer Review

Peer Review: Die grundsätzliche Bedeutung einer Webseite

Im Ansatz des „**Peer Review**“ wird die folgende Annahme gemacht:

Wenn eine Webseite i einen Link auf eine Webseite j enthält, dann

1. gibt es eine inhaltliche Beziehung zwischen beiden Webseiten, und
2. der Autor der Webseite i hält die Informationen auf Webseite j für wertvoll.

Die Link-Struktur des Internets, also der **Webgraph**, spielt im Peer-Review eine besondere Rolle. Zur Erinnerung:

- Die Webseiten sind Knoten und
- die Hyperlinks sind die gerichteten Kanten des Webgraphen.

Es gibt verschiedene Peer-Review Verfahren, beispielsweise

das von *Google* genutzte **Page-Rank** Verfahren von Brin und Page

oder das HITS-Verfahren,

Hypertext Induced Topic Search von Kleinberg.

Page-Rank: Notation

Um die „grundlegende Bedeutung“ einer Webseite zu messen, berücksichtigt der Page-Rank nur die Link-Struktur des Internets, nicht aber den Inhalt der Seite.

Wir schreiben im Folgenden $G = (V, E)$, um den Web-Graphen zu bezeichnen.

- Der Einfachheit halber nehmen wir an, dass die Webseiten mit den Zahlen $1, \ldots, n$ durchnummeriert sind, und dass $V = \{1, 2, \ldots, n\}$ gilt.
- Für jeden Knoten $i \in V$ ist

$$a_i \;:=\; \text{Aus-Grad}_G(i)$$

  der Ausgangsgrad von i in G, also die Anzahl der Hyperlinks, die von der Webseite i auf andere Webseiten verweisen.
- Für eine Webseite $j \in V$ schreiben wir $\text{Vor}_G(j)$, um die Menge aller Webseiten zu bezeichnen, die einen Link auf j enthalten, d.h.

$$\text{Vor}_G(j) \;=\; \{i \in V \;:\; (i,j) \in E\}.$$

  Die Elemente in $\text{Vor}_G(j)$ heißen Vorgänger von j.

Page-Rank mittels Peer Review

Wir messen die „grundlegende Bedeutung“ einer Webseite i durch die Zahl $\mathbf{PR_i}$, den Page-Rank von i.

- Der Wert PR_i soll die Qualität, im Sinne von „Renommee“ oder „Ansehen“, der Webseite i widerspiegeln;
- die Zahl PR_i soll umso größer sein, je höher das Renommee der Webseite i ist.

Wann sollte Webseite i hoch bewertet werden?

*Wenn genügend viele **hochbewertete** Webseiten auf i zeigen!*

Nehmen wir doch einfach mal an, dass wir alle Page-Ranks PR_i bestimmt haben.

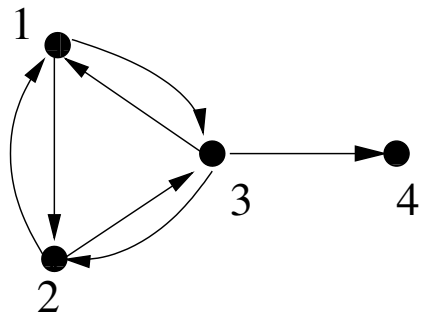
1. Wir fordern, dass eine Webseite i ihren Page-Rank an alle Webseiten j zu gleichen Maß*en* „vererbt“, auf die i zeigt, für die also $(i,j) \in E$ gilt.
2. Mit dieser Sichtweise müsste also für alle $j \in V$ mit $\mathrm{Vor}_G(j) \neq \emptyset$ gelten:

$$PR_j = \sum_{i \in \mathrm{Vor}_G(j)} \frac{PR_i}{a_i}.$$

Schauen wir mal, was passiert

Senken, also Knoten vom Ausgangsgrad 0, vererben ihren Page-Rank nicht.
Ist das problematisch?

Betrachte den folgenden „Webgraphen“ $G = (V, E)$:



Die einzigen Page-Rank Werte, die die Gleichungen

$$\mathrm{PR}_j = \sum_{i \in \mathrm{Vor}_G(j)} \frac{\mathrm{PR}_i}{a_i}.$$

erfüllen, sind $\mathrm{PR}_1 = \mathrm{PR}_2 = \mathrm{PR}_3 = \mathrm{PR}_4 = 0$ und diese Werte sollen die

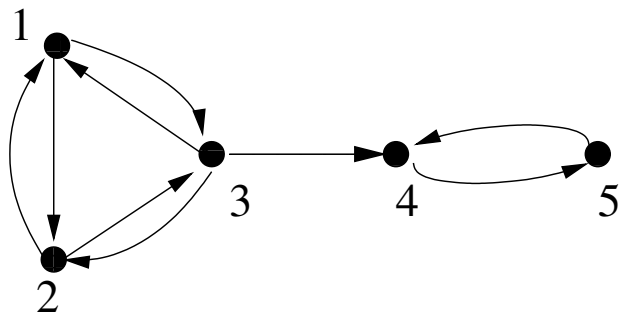
„grundlegende Bedeutung“

der 4 Seiten widerspiegeln?

Und damit nicht genug

Ein weiteres Problem stellen Knoten dar, die zwar unter sich verbunden sind, die aber keine Kante zu einem anderen Knoten des Graphen G enthalten.

Betrachten den folgenden Graphen $G = (V, E)$:



Man kann sich leicht davon überzeugen, dass Page-Rank Werte die Gleichung

$$\mathrm{PR}_j = \sum_{i \in \mathrm{Vor}_G(j)} \frac{\mathrm{PR}_i}{a_i}.$$

genau dann erfüllen, wenn $\mathrm{PR}_1 = \mathrm{PR}_2 = \mathrm{PR}_3 = 0$ und $\mathrm{PR}_4 = \mathrm{PR}_5$ gilt.

Das darf doch wohl nicht wahr sein!

Gehört unser Ansatz „in die Tonne“
oder stimmt die Grundidee?

Woran liegt's?

In allen schlechten Beispielen waren die Graphen

nicht stark zusammenhängend!.

(a) Der Ausweg?

- Füge Kanten von einer Webseite i zu **allen** anderen Webseiten ein,
- „**dämpfe**" aber den Beitrag der neuen Seiten mit dem Faktor $1 - d$ für $0 \leq d \leq 1$.

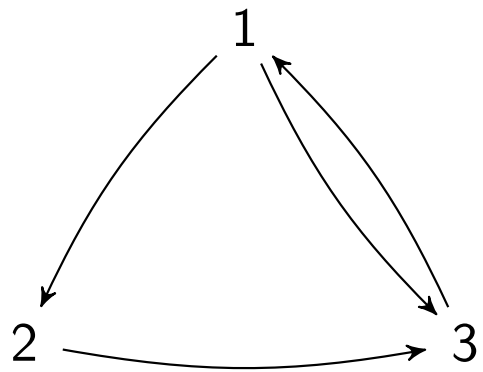
(b) Ein Tupel $\mathrm{PR} = (\mathrm{PR}_1, \ldots, \mathrm{PR}_n) \in \mathbb{R}^n$ hat die **Page-Rank Eigenschaft** bezüglich d, wenn für alle $j \in V$ gilt:

$$\mathrm{PR}_j \quad = \quad \frac{1-d}{n} \quad + \; d \cdot \sum_{i \in \mathrm{Vor}_G(j)} \frac{\mathrm{PR}_i}{a_i}.$$

Der neue Page-Rank: Ein Beispiel

Wir setzen $d := \frac{1}{2}$.

Betrachte den folgenden Graphen $G = (V, E)$:



Wir suchen ein Tupel $\mathrm{PR} = (\mathrm{PR}_1, \mathrm{PR}_2, \mathrm{PR}_3)$ von reellen Zahlen, so dass gilt:

(1) $\mathrm{PR}_1 = \frac{1}{2 \cdot 3} + \frac{1}{2} \cdot \frac{\mathrm{PR}_3}{1}$,

(2) $\mathrm{PR}_2 = \frac{1}{2 \cdot 3} + \frac{1}{2} \cdot \frac{\mathrm{PR}_1}{2}$

(3) $\mathrm{PR}_3 = \frac{1}{2 \cdot 3} + \frac{1}{2} \cdot \left(\frac{\mathrm{PR}_1}{2} + \frac{\mathrm{PR}_2}{1} \right)$.

Wir lösen das lineare Gleichungssystem und erhalten

$$\mathrm{PR}_1 = \frac{14}{39}, \quad \mathrm{PR}_2 = \frac{10}{39}, \quad \mathrm{PR}_3 = \frac{15}{39}.$$

Page-Rank: Die wichtigen Fragen

Wir müssen auch im Allgemeinen ein lineares Gleichungssystem lösen.

(a) Ist das Gleichungssystem überhaupt lösbar und wenn ja, ist die Lösung eindeutig?

(b) Und wie soll man, bitte schön,
ein Gleichungssystem mit mehreren Milliarden Zeilen und Spalten lösen?

- Unsere Rechner sind mittlerweile so mächtig: kein Problem mit Gaußscher Eliminierung!
- **Denkste!** Ein Gleichungssystem dieser Dimension können wir auch mit allen Rechnern dieser Welt nicht knacken, wenn ...
  ...wir eine Gaußsche Eliminierung ausführen müssen.

(c) Und selbst wenn es genau eine Lösung PR gibt und wir diese Lösung irgendwie bestimmen können:

Gibt PR_i das Renommee der Webseite i wieder?

Page-Rank mittels Zufalls-Surfer

Verteilungen

Ein Tupel $\pi \in \mathbb{R}^n$ heißt eine **Verteilung** (auf $\{1, \ldots, n\}$), falls π die beiden folgenden Eigenschaften hat:

- $\pi_i \geqslant 0$ für alle $i \in \{1, \ldots, n\}$ und
- $\sum_{i=1}^{n} \pi_i = 1$ gilt.

Wir benutzen Verteilungen, um **Irrfahrten** (engl: Random Walks) in einem gerichteten Graphen $G = (V, E)$ zu beschreiben:

Dazu legen wir für jeden Knoten $k \in V$ eine Verteilung π^k fest, so dass π^k_i die Wahrscheinlichkeit ist, in einem Schritt von k zum Knoten i zu springen.

Page-Rank mittels Zufalls-Surfer

Warum nicht PR_j als Wahrscheinlichkeit definieren, dass

Seite j von einem „**zufällig** im Web herumirrenden Surfer“ besucht wird?

- Lass den Surfer auf einer zufällig ausgewürfelten Seite i für k Schritte laufen und bestimme die Wahrscheinlichkeit

$$\pi_{i,j}^{(k)}$$

  mit der Seite j im $k$ten Schritt besucht wird. Bewerte j mit $\frac{1}{n} \cdot \sum_{i=1}^{n} \pi_{i,j}^{(k)}$.
- Aber welchen Wert von k sollten wir nehmen? Definiere

$$\mathrm{PR}_j^* = \lim_{k \to \infty} \frac{1}{n} \cdot \sum_{i=1}^{n} \pi_{i,j}^{(k)}$$

  als Page-Rank „aus Sicht des Zufalls-Surfers“!

Und mit welcher Wahrscheinlichkeit soll unser Surfer von i auf j springen?

Die Übergangsmatriz $P_d(G)$

Übergangswahrscheinlichkeiten

Mit welcher Wahrscheinlichkeit $P_d(G)_{i,j}$ soll unser Surfer von i auf j springen?

Wir definieren die Übergangswahrscheinlichkeit $P_d(G)_{i,j}$ für $i, j \in \{1, \ldots, n\}$ durch

$$P_d(G)_{i,j} = \begin{cases} \frac{1-d}{n} + \frac{d}{a_i} & \text{falls } (i,j) \text{ eine Kante von } G \text{ ist,} \\ \frac{1-d}{n} & \text{sonst.} \end{cases}$$

(G ist der Webgraph, d der Dämpfungsfaktor und $a_i = \text{Aus-Grad}_G(i)$.)

Wir haben die „**Übergangsmatrix**“, also die Matrix der Übergangswahrscheinlichkeiten definiert. Diese Matrizen heißen auch stochastische Matrizen:

Eine $n \times n$ Matrix P heißt **stochastisch**, wenn

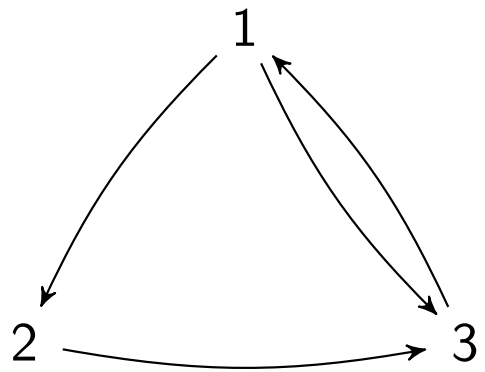
(1) $P_{i,j} \geqslant 0$ für alle $i, j \in \{1, \ldots, n\}$, und

(2) für jede Zeile $i \in \{1, \ldots, n\}$ gilt: $\sum_{j=1}^{n} P_{i,j} = 1$.

Die Matrix P ist also genau dann stochastisch, wenn jede Zeile eine Verteilung ist.

Die Übergangsmatrix

Für den Wert $d = \frac{1}{2}$ und den Graphen G



ist beispielsweise $P_{1/2}(G)_{1,1} = \frac{1}{6}$, $P_{1/2}(G)_{1,2} = \frac{1}{6} + \frac{1}{4} = \frac{5}{12}$ und $P_{1/2}(G)_{2,3} = \frac{1}{6} + \frac{1}{2} = \frac{2}{3}$. Die vollständige Übergangsmatrix ist

$$P_{1/2}(G) = \begin{pmatrix} \frac{1}{6} & \frac{5}{12} & \frac{5}{12} \\ \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \\ \frac{2}{3} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}.$$

Peer Review und Zufalls-Surfers

Wie passen der Page-Rank **PR** (definiert durch „Peer Review“) und der Page-Rank $\mathbf{PR}^*$ (definiert durch „Zufalls-Surfer“) zusammen?

- Wenn die beiden Sichtweisen verschiedene Bewertungen ergeben, welche sollen wir dann wählen?
- Können wir denn zumindest alle Page-Ranks PR_i^* scharf approximieren?

Wir können diese Fragen mit Hilfe von Markov-Ketten beantworten!

Markov-Ketten

Markov-Ketten und Grenzwahrscheinlichkeiten

$G = (V, E)$ sei ein gerichteter Graph mit Knotenmenge $V = \{1, \ldots, n\}$.

Eine (homogene) **Markov-Kette** wird durch das Paar (G, P) beschrieben.

(a) G hat keine Senke, d.h. Aus-Grad$_G(v) > 0$ gilt für alle Knoten v von G.

(b) Die Matrix P ist eine stochastisch Matrix mit n Zeilen und n Spalten. Es ist $P_{i,j} = 0$ genau dann, wenn (i, j) keine Kante in G ist.

Man nennt G den **Graph** der Kette und P ihre **Übergangsmatrix**, die Knoten von G nennt man auch **Zustände**.

- Was tut eine Markov-Kette (G, P)?
  - Sie definiert eine Irrfahrt, in der ein „Zufalls-Surfer“ mit Wahrscheinlichkeit $P_{i,j}$ vom Knoten i zum Knoten j springt.
- Bestimme die **Grenzwahrscheinlichkeiten** $\pi_{i,j}$. Zur Erinnerung:
  - $\pi_{i,j}^{(k)}$ ist die Wahrscheinlichkeit, dass eine im Knoten i beginnende Irrfahrt nach k Schritten im Knoten j endet und es ist
  - $$\pi_{i,j} = \lim_{k \to \infty} \pi_{i,j}^{k}.$$

„Schöne“ Markov-Ketten

Eine Markov-Kette (G, P) mit $G = (V, E)$ ist **ergodisch**, wenn

(a) die Grenzwahrscheinlichkeiten $\pi_{i,j}$ und $\pi_{i',j}$ für alle Knoten i, i' und j existieren und übereinstimmen sowie

(b) $\pi_{i,j} > 0$ für alle Knoten i, j gilt.

In **Mathe 3** wird gezeigt:

Eine Markoff-Kette (G, P) ist genau dann ergodisch, wenn

G **stark zusammenhängend** und **aperiodisch** ist.

G ist genau dann aperiodisch, wenn für alle Knoten i Eins der größte gemeinsame Teiler aller Weglängen von i nach i ist.

PR^* und Grenzverteilung stimmen überein!

1. $G = (V, E)$ mit $V = \{1, \ldots, n\}$ ist der Webgraph und d sei der Dämpfungsfaktor.
2. Wir erinnern an die Übergangswahrscheinlichkeit $P_d(G)_{i,j}$ für $i, j \in \{1, \ldots, n\}$

$$P_d(G)_{i,j} = \begin{cases} \frac{1-d}{n} + \frac{d}{a_i} & \text{falls } (i,j) \text{ eine Kante von } G \text{ ist,} \\ \frac{1-d}{n} & \text{sonst.} \end{cases}$$

3. Welche Markov-Kette modelliert die Irrfahrten des Zufalls-Surfers?
   - Sei $\vec{K}_n = (V, E_n)$ der **vollständige, gerichtete Graph** mit Knotenmenge $V = \{1, \ldots, n\}$ und Kantenmenge $E_n = \{ (u, v) \ : \ u, v \in \{1, \ldots, n\}, \ u \neq v \}$.
   - Wir nennen $(\vec{K}_n, P_d(G))$ die **Webkette**.
   - Die Webkette ist ergodisch: $\pi_{i,j} = \pi_{i',j}$ gilt für alle Knoten i, i' und j.
     Also ist $\lim_{k\to\infty} \frac{1}{n} \sum_{i=1}^{n} \pi_{i,j}^{(k)} = \frac{1}{n} \cdot \sum_{i=1}^{n} \lim_{k\to\infty} \pi_{i,j}^{(k)} = \frac{1}{n} \cdot \sum_{i=1}^{n} \pi_{i,j} = \pi_{1,j}$.

Für die „**Grenzverteilung**“ $\rho = (\pi_{1,1}, \ldots, \pi_{1,n})$ gilt

$$PR^* = \rho.$$

Markov-Ketten: Was müssen wir uns anschauen?

(a) Die Bewertung PR^* aus Sicht des Zufalls-Surfers stimmt überein mit der

Grenzverteilung ρ

der Webkette.

(b) Entspricht auch PR, also die Bewertung aus Sicht des Peer Reviews, einem fundamentalen Begriff aus der Theorie der Markov-Ketten?

Wir beginnen mit einem Einschub über die Matrixmultiplikation.

Matrizenprodukt und Matrix-Vektor Produkt

A, B seien $n \times n$ Matrizen reeller Zahlen und $x \in \mathbb{R}^n$ sei ein Tupel reeller Zahlen.

- Das **Matrizenprodukt**: Um den Eintrag $(A \cdot B)_{i,j}$ der Produktmatrix $A \cdot B$ zu bestimmen, multipliziere die $i$te Zeile von A mit der j Spalte von B, also

$$(A \cdot B)_{i,j} = \sum_{k=1}^{n} A_{i,k} \cdot B_{k,j}.$$

- **Matrix-Vektor Produkte**:
  - Um $y = x^T \cdot A$ zu bestimmen, interpretiere x als Zeilenvektor, den man dann nacheinander mit allen Spalten von A multiplizieren muss, also

$$y_i = \sum_{k=1}^{n} x_k \cdot A_{k,i}.$$

  - Um $z = A \cdot x$ zu bestimmen, interpretiere x als Spaltenvektor, den man dann nacheinander mit allen Zeilen von A multiplizieren muss, also

$$z_i = \sum_{k=1}^{n} A_{i,k} \cdot x_k.$$

Ein Schritt einer Markov-Kette

WOOOOOOW (1/2)

Wieso reden wir plötzlich über Matrizen? Und wenn wir schon mal dabei sind: Warum werden Matrizenprodukte und Matrix-Vektor Produkte so definiert?

Sei (G, P) eine Markov-Kette. Wenn wir mit Wahrscheinlichkeit π_i im Knoten i starten, dann sind wir nach einem Schritt im Knoten j mit Wahrscheinlichkeit

$$\sum_{i=1}^{n} \pi_i \cdot P_{i,j} \overset{\text{toll!}}{=} (\pi^T \cdot P)_j.$$

Die Kette,

wenn in Verteilung π gestartet,

befindet sich nach einem Schritt in der Verteilung

$$\pi^T \cdot P.$$

Und wenn wir die Markov-Kette zwei Schritte lang beobachten?

WOOOOOOW (2/2)

Wir wählen einen Startzustand i der Kette (G, P) mit Wahrscheinlichkeit π_i.

1. Nach dem ersten Schritt ist die Kette im Zustand ℓ mit Wahrscheinlichkeit
$$\sum_{i=1}^{n} \pi_i \cdot P_{i,\ell} = (\pi^T \cdot P)_\ell.$$
2. Nach dem 2. Schritt ist die Kette im Zustand j mit Wahrscheinlichkeit
$$\sum_{\ell=1}^{n} (\pi^T \cdot P)_\ell \cdot P_{\ell,j} = \sum_{\ell=1}^{n} \Big(\sum_{i=1}^{n} \pi_i \cdot P_{i,\ell} \Big) \cdot P_{\ell,j} = \sum_{i=1}^{n} \pi_i \cdot \Big(\sum_{\ell=1}^{n} P_{i,\ell} \cdot P_{\ell,j} \Big)$$
$$\overset{\text{toll!}}{=} \sum_{i=1}^{n} \pi_i \cdot P^2_{i,j} = \left(\pi^T \cdot P^2\right)_j.$$

Nach k Schritten befindet sich die Kette im Zustand j mit Wahrscheinlichkeit

$$\left(\pi^T \cdot P^k\right)_j.$$

Stationäre Verteilungen

Sei (G, P) eine Markov-Kette und es gelte $V = \{1, \ldots, n\}$. Eine Verteilung π auf V heißt

stationär für die Markov-Kette (G, P),

falls die Kette „nach einem Schritt in π verbleibt, wenn sie in π gestartet wird“.

Was genau bedeutet die Eigenschaft „stationär zu sein“?

- Wir wissen bereits, dass die Kette (G, P), wenn in der Verteilung π gestartet, sich nach einem Schritt in der Verteilung $\pi^T \cdot P$ befindet.
- Die Verteilung π ist also genau dann stationär, wenn gilt

$$\pi^T \cdot P = \pi.$$

In der **Mathe 3** wird gezeigt:

Sei (G, P) eine ergodische Kette.

(a) Dann besitzt die Kette genau eine stationäre Verteilung σ und

(b) σ stimmt überein mit der Grenzverteilung $\rho = (\pi_{1,1}, \ldots, \pi_{1,n})$.

PR und stationäre Verteilungen stimmen überein!

Wir fordern, dass der Pange-Rank PR eine Verteilung ist.

Wir haben den Page-Rank Vektor PR definiert durch

$$\begin{aligned} \mathrm{PR}_j &= \frac{1-d}{n} + d \cdot \sum_{i \in \mathrm{Vor}_G(j)} \frac{\mathrm{PR}_i}{a_i} \\ &\overset{!}{=} \sum_{i \in \{1,\ldots,n\}} \frac{1-d}{n} \cdot \mathrm{PR}_i + d \cdot \sum_{i \in \mathrm{Vor}_G(j)} \frac{\mathrm{PR}_i}{a_i} \\ &= \sum_{i \in \{1,\ldots,n\}} \mathrm{PR}_i \cdot P_n[i,j] = (\mathrm{PR}^T \cdot P_n)_j. \end{aligned}$$

Wenn der Page-Rank eine Verteilung ist:

(a) Der Page-Rank ist eine stationären Verteilung der Webkette und

(b) stimmt überein mit der Grenzverteilung.

Peer Review $\equiv$ Zufalls-Surfer

Wir haben den Page-Rank Vektor auf zwei Arten definiert, nämlich

(a) über den Peer Review als stationäre Verteilung PR π_S der Webkette und

(b) über den Zufalls-Surfer als Grenzverteilung PR* der Webkette.

Es ist PR $=$ PR*: Die beiden Page-Rank Definitionen (über Peer Review bzw. über Zufalls-Surfer) stimmen überein.

Wenn wir jetzt noch den Page-Rank mit vertretbarem Aufwand berechnen könnten, ist alles gut :-)))

Page-Rank: Eine effiziente Approximation (1/2)

Es genügt, die Grenzverteilung ρ der Webkette $(\vec{K_n}, P_d(G))$ effizient zu approximieren, denn Grenzverteilung und stationäre Verteilung stimmen überein.

1. Die Grenzverteilung ρ ist unabhängig von der Anfangsverteilung.
   Sei π^0 eine beliebige Verteilung auf $\{1, \ldots, n\}$.
2. Wir wissen, dass sich die Kette nach einem Schritt in der Verteilung $\pi^1 := (\pi^0)^T \cdot P$ befindet.
3. Mit vollständiger Induktion folgt, dass sich die Kette nach $k+1$ Schritten in der Verteilung

$$\pi^{k+1} := (\pi^k)^T \cdot P$$

befindet und nach Definition der Grenzverteilung gilt

$$\rho = \lim_{k \to \infty} \pi^k.$$

Page-Rank: Eine effiziente Approximation (2/2)

Es ist

$$\rho = \lim_{k \to \infty} \pi^k,$$

mit einer beliebigen Verteilung π^0 und der Rekursion

$$\pi^{k+1} := \pi^k \cdot P.$$

Approximiere ρ durch π^k für ein „genügend großes“ k.

- ✓ Die Berechnung des Matrix-Vektor Produkts $\pi^k \cdot P$ ist hochgradig parallelisierbar.
- ✓ Es ist $\rho \approx \pi^k$ bereits für kleine Werte von k:
  - Dies folgt vor allem aus der Tatsache, dass das Web „hoch-gradig zusammenhängend“ ist: Das „**small-world Phänomen**“ besagt z.B, dass die durchschnittliche Distanz zwischen zwei Webseiten sehr klein ist.

:-) Es ist alles gut!

Zusammenfassung

Zusammenfassung: Homogene Markov-Ketten

Eine Markov-Kette (G, P) besteht aus einer stochastischen Übergangsmatrix P und einem gerichteten Graphen $G = (\{1, \ldots, n\}, E)$ mit $(i, j) \in E \iff P_{i,j} > 0$.

(a) Ein Schritt der Markov-Kette (G, P) kann durch das Matrix-Vektor Produkt $\pi \cdot P$ beschrieben werden:

- Wenn ein Zustand i mit Wahrscheinlichkeit π_i ausgewürfelt wird,
- dann befindet sich die Kette nach einem Schritt im Zustand j mit Wahrscheinlichkeit $(\pi^T \cdot P)_j$.

(b) (G, P) ist **ergodisch**, wenn G stark zusammenhängend und aperiodisch ist.

- Eine ergodische Kette besitzt genau eine **stationäre Verteilung** ρ
- und ρ ist die **Grenzverteilung**

  d.h. die Wahrscheinlichkeit, dass Zustand i am Ende einer einer genügend langen Irrfahrt angenommen wird, konvergiert gegen ρ_i.

(c) Eine stationäre Verteilung ρ ist Lösung des linearen Gleichungssystems

$$\rho^T \cdot P = \rho.$$

Zusammenfassung: Anwendungen von Markov-Ketten

(a) Modelliere ein **Glückspiel** (mit 1 € Einsatz und 1 € Gewinn/Verlust):
- Verwende die Zustände $0, \ldots, n$ und Übergänge vom Zustand $0 < i < n$ zu den Zuständen $\max\{0, i-1\}$ und $\min\{i+1, n\}$ mit Wahrscheinlichkeit $1/2$.
- Der Spieler ist im Zustand 0 und die Bank im Zustand n ruiniert: Die Übergänge von i nach i haben für $i = 0$ und $i = n$ die Wahrscheinlichkeit 1.

(b) Modelliere eine **Warteschlange** an einer Supermarktkasse:
- Verwende die Zustände $0, \ldots, n$ und Übergänge vom Zustand i zu den Zuständen $\max\{0, i-1\}, i, \ldots, n$.
- Die Wahrscheinlichkeit, dass sich eine Schlange verlängert, wächst mit der Länge der Schlange. Im besten Fall reduziert sich die Länge um Eins.

(c) Varianten von Markov-Ketten sind:
- Ketten mit unbeschränkt vielen Zuständen.
- **Inhomogene Ketten**, in denen die Wahrscheinlichkeit eines Übergangs von einem Zustand i zu einem Zustand j nicht nur von i und j, sondern auch vom Zeitpunkt des Übergangs abhängt.
- Ketten **kter Ordnung**, in denen die Wahrscheinlichkeit des neuen Zustands von den k letzten Zuständen abhängt.

(d) Weitere Anwendungen in der Bioinformatik (Auffinden von CpG Inseln), Finanzmathematik (Modellierung von Aktienkurs- und Zinsentwicklungen), im Entwurf und in der Analyse von Algorithmen, ... und der Page-Rank.

Zusammenfassung: Page-Rank

Für eine Anfrage a wählt Google zuerst eine Menge $\mathcal{M}(a)$ von Webseiten aus, die für a relevant sind. Die Seiten werden dann nach ihrem Page-Rank geordnet.

(a) Der Ansatz des **Peer-Reviews**:

- Die Webseite v erhält anteilig das Renommee einer jeden Seite u, die einen Hyperlink auf v gesetzt hat.
- v vererbt ihr Renommee anteilig auf jede Webseite w, auf die sie einen Hyperlink gesetzt hat.

Die Berechnung des Page-Ranks führt auf ein lineares Gleichungssystem.

(b) Der Ansatz des **Zufalls-Surfers**:

- Der Page-Rank der Seite v ist die Wahrscheinlichkeit, dass v am Ende einer hinreichend langen Irrfahrt besucht wird.
- Der Page-Rank wird bereits für kurze Irrfahrten scharf approximiert.

(c) Durch das Hinzufügen von „neuen Kanten“ wird die Webkette **ergodisch**: Die Grenzverteilung ist die einzige stationäre Verteilung.

- Die **Grenzverteilung** stimmt mit dem Page-Rank aus Sicht des Peer-Reviews
- und die **stationäre Veteilung** mit dem Page-Rank aus Sicht der Zufalls-Surfers überein.