Die Größe des Netzes

- Schätzungen gehen weit auseinander:
  - Über eine Milliarde im Gebrauch befindliche IP-Adressen
  - Zwischen 20 Milliarden und einer Billion indizierte Webseiten.
- Ungefähr 200 Millionen Websites und mehr als 1.5 Milliarden Internetnutzer.
- Die Gesamtgröße des indizierten Webs liegt im Bereich von mehreren Hundert Terabytes.

  Die durchschnittliche Größe eines Dokuments liegt zwischen 5 und 10 Kilobytes.
- Im .com Bereich scheinen sich 40% aller Dokumente täglich zu ändern; rund die Hälfte aller Webseiten scheint eine Lebenszeit von nur 10 Tagen zu besitzen.

Suchmaschinen:
Für einen sich rasant ändernden Suchraum gigantischer Größe sind Anfragen ohne merkliche Reaktionszeit zu beantworten.

Der Aufbau einer Suchmaschine

Der Webgraph besteht aus den durch Hyperlinks verbundenen Webdokumenten.

1. Der Crawler traversiert den **Webgraphen** mit Hilfe der Hyperlinks: Die gefundenen Dokumente werden erfasst (und damit indiziert).
2. Die Dokumente werden in einem **Web Repository** verwaltet.
   - Das **Einfügen** und **Entfernen** von Webseiten ist zu unterstützen.
   - Dokumente zu einer gegebenen Stichwortmenge müssen schnell **auffindbar** sein.
   - Man verwendet unter anderem B-Bäume und Hashing.
3. Die gefundenen Dokumente für eine Stichwortmenge sind zu bewerten. Wie?
   - Syntaktische Eigenschaften wie Häufigkeit der Stichworte, ihre Schriftgröße werden ausgenutzt
   - ebenso wie das Vorkommen der Stichworte in Hyperlinks, die auf die Weibseite zeigen sowie
   - eine **Bewertung der Relevanz** der Webseite.

Connectivity Analyse und Page Rank

Connectivity Analyse

Die zentrale Annahme der Connectivity Analyse: Wenn ein Dokument *A* auf Dokument *B* zeigt, dann

- gibt es eine inhaltliche Beziehung zwischen den beiden Dokumenten und
- der Autor des Dokuments *A* hält Dokument *B* für wertvoll.

Und die Konsequenz:

- Schließe von der relativen Wertschätzung zwischen Dokumenten auf die absolute Relevanz der Dokumente.
- Bewerte Dokumente aufgrund der Graphstruktur des WWW.

Relevanz Bewertung durch Google

- Syntaktische Eigenschaften der Stichworte werden benutzt.
- Der **Page-Rank** $\mathtt{pr}(w)$ einer Webseite w ist die zentrale Komponente in der Bewertung.

  Bewerte die Relevanz der Webseite ohne Bezug auf Stichworte.

- Verschiedene Sichtweisen in der Berechnung des Page-Rank:
  - Der **Random Surfer**
    - ★ springt zufällig von Seite zu Seite.
    - ★ $\mathtt{pr}(w)$ ist die relative Häufigkeit des Besuchs der Seite w.
  - Im **Peer-Review** wird angenommen, dass ein Dokument u mit d_u Hyperlinks seinen Page-Rank $\mathtt{pr}(u)$ gleichmäßig verteilt:
    - ★ Wenn u auf Dokument w zeigt, dann „erbt“ w den Bruchteil $\frac{\mathtt{pr}(u)}{d_u}$.
    - ★ Die Seite w erbt insgesamt den Betrag

$$\mathrm{pr}(w) = \sum_{u \text{ zeigt auf w}} \frac{\mathrm{pr}(u)}{d_u}.$$

- Verschiedene Sichtweisen, identische Resultate?

Markoff-Ketten: Die stationäre Verteilung

Der Random Surfer und Markoff-Ketten

Eine Matrix P heißt **stochastisch**, wenn

$$P \geqslant 0 \text{ und } \sum_w P[v, w] = 1 \text{ für alle Zeilen } v \text{ gilt.}$$

Eine **Markoff-Kette** (G, P) besteht aus

- einem gerichteten Graphen $G = (V, E)$ und
- einer stochastischen Übergangsmatrix P, deren Zeilen und Spalten mit Knoten aus V indiziert sind.

- Eine Markoff-Kette (G, P) definiert einen Random Walk auf G:
  - $P[v, w]$ ist die Wahrscheinlichkeit, dass der Random Walk zum Knoten w wechselt, wenn Knoten v erreicht wird.
- Unsere Anwendung:
  - Wähle G als den Webgraphen und $P[v, w] = \frac{1}{d_v}$, wenn die Webseite v auf die Webseite w zeigt und v genau d_v Hyperlinks besitzt.
  - Der Random Walk modelliert den Random Surfer.

Peer-Review und die stationäre Verteilung

Sei (G, P) eine Markoff-Kette mit $G = (V, E)$. Die Verteilung $\pi = (\pi_v \mid v \in V)$ heißt **stationär**, falls $\pi^T = \pi^T \cdot P$ gilt.

- Wählen wir einen Startknoten gemäß einer stationären Verteilungen π und führen einen Schritt der Markoff-Kette durch, dann verbleiben wir in π.
- Wenn der Page-Rank nach der Perspektive des Peer-Review definiert wird, dann ist

$$\mathrm{pr}(w) = \sum_{u \text{ zeigt auf w}} \frac{\mathrm{pr}(u)}{d_u}.$$

- Wenn die Übergangsmatrix P durch

$$P[u, w] = \begin{cases} \frac{1}{d_u} & (u, w) \text{ ist ein Hyperlink} \\ 0 & \text{sonst.} \end{cases}$$

definiert wird, dann ist der Page-Rank eine stationäre Verteilung!

Random Surfer und Grenzverteilung (1/3)

Sei $p_{i,j}^{(k)}$ die Wahrscheinlichkeit den Knoten j vom Knoten i aus in k Schritten zu erreichen. Dann ist

$$p_{i,j}^{(k)} = P^k[i,j].$$

- Induktion nach k: Die Verankerung für $k = 1$ ist offensichtlich.
- Nach Induktionsannahme gilt

$$p_{i,j}^{(k+1)} = \sum_r p_{i,r}^{(k)} \cdot P[r,j] = \sum_r P^k[i,r] \cdot P[r,j] = P^{k+1}[i,j].$$

und das war zu zeigen.

Die Grenzwahrscheinlichkeit $\lim_{k\to\infty} p_{i,j}^{(k)}$, den Knoten j vom Knoten i aus zu erreichen, stimmt überein mit

$$\lim_{k\to\infty} P^k[i,j].$$

Es ist gefährlich, wenn die Grenzverteilung $\lim_{k\to\infty} P^k[i,j]$ vom Startknoten i abhängt.

Die Markoff-Kette (G,P) heißt genau dann **ergodisch**, wenn für alle i_1, i_2 und j

$$\lim_{k\to\infty} P^k[i_1,j] = \lim_{k\to\infty} P^k[i_2,j] > 0 \text{ gilt.}$$

- Eine Markoff-Kette (G,P) ist genau ergodisch, wenn
  - G **irreduzibel** ist: Zu je zwei Knoten i und j gibt es einen Weg von i nach j
  - und G **aperiodisch** ist: Es gibt kein $d > 1$, so dass die Länge aller Wege, die zu ihrem Startknoten zurückkehren, durch d teilbar ist.
- Der Webgraph ist sicherlich aperiodisch, aber nicht irreduzibel: Webseiten ohne Hyperlinks sind Sackgassen!

Random Surfer und Grenzverteilung (3/3)

Google legt deshalb einen Random Walk zugrunde, der

- mit Wahrscheinlichkeit $(1 - d)$ ein benachbartes Dokument aufsucht und
- mit Wahrscheinlichkeit d zu einem zufälligen Dokument springt.

- Wenn wir annehmen, dass es genau n Dokumente gibt, dann erhalten wir die Übergangsmatrix

$$P'[u, w] = \begin{cases} (1 - \frac{d}{n}) \cdot \frac{1}{d_u} & (u, w) \text{ ist ein Hyperlink} \\ \frac{d}{n} & \text{sonst.} \end{cases}$$

- Die neue „Web-Kette“ ist aperiodisch und irreduzibel.
- Die **Random Surfer Perspektive**: Der Page-Rank stimmt mit der Grenzverteilung überein und die Grenzverteilung hängt nicht von der Anfangsverteilung ab.

Peer-Review = Random Surfer

Jede ergodische Markoff-Kette (G, P) besitzt genau eine stationäre Verteilung, nämlich die Grenzverteilung π^∞ mit

$$\pi_j^\infty = \lim_{k \to \infty} P^k[1, j].$$

- Für jede Verteilung π ist $\pi^T \cdot P^\infty = (\pi^\infty)^T$:

$$\begin{aligned}
\pi^T \cdot P^\infty &= \left(\sum_{i=1}^{N} \pi_i \cdot P^\infty[i,1], \ \ldots, \ \sum_{i=1}^{N} \pi_i \cdot P^\infty[i,N] \right) \\
&= \left(P^\infty[1,1] \cdot \sum_{i=1}^{N} \pi_i, \ \ldots, \ P^\infty[1,N] \cdot \sum_{i=1}^{N} \pi_i \right) \\
&= (P^\infty[1,j] \mid j = 1, \ldots, N) = (\pi^\infty)^T.
\end{aligned}$$

- Ist π^∞ wirklich stationär?

Stationäre Verteilungen und ergodische Ketten

- Die Grenzverteilung π^∞ ist stationär, denn

$$\begin{aligned}(\pi^\infty)^T &= (\pi^\infty)^T \cdot P^\infty = (\pi^\infty)^T \cdot \lim_{k\to\infty} P^k \\ &= (\pi^\infty)^T \cdot \lim_{k\to\infty} P^{k+1} = (\pi^\infty)^T \cdot P^\infty \cdot P = (\pi^\infty)^T \cdot P.\end{aligned}$$

- Sei π eine beliebige stationäre Verteilung.
  - Es ist $\pi^T = \pi^T \cdot P$ und damit natürlich auch $\pi^T = \pi^T \cdot P^\infty$.
  - Andererseits wissen wir: $\pi^T \cdot P^\infty = (\pi^\infty)^T$.
  - Die Eindeutigkeit der stationären Verteilung ist nachgewiesen.

Die Perspektiven des Peer Review und des Random Surfers sind identisch!

Die Berechnung des Page-Rank

Sollten wir versuchen, das lineare Gleichungssystem $\pi^T \cdot P' = \pi$ zu lösen?

- Sehr **naiver** Versuch bei einer Matrix P' mit hunderten Milliarden Zeilen und Spalten!

Es ist $\lim_{k\to\infty} \pi_0^T \cdot (P')^k = \mathrm{pr}$ für jede Anfangsverteilung π_0.

- Beginne mit der uniformen Verteilung π_0 und führe die Iteration $\pi_{k+1} = \pi_k^T \cdot P'$ aus.
- Wir haben Glück:
  - Schnelle Konvergenz: Das Netz ist **hochgradig zusammenhängend**.
  - Relativ wenige Hyperlinks pro Seite $\Rightarrow$ Die Übergangsmatrix P' ist **dünn besetzt**.
  - Die Berechnung des **Matrix-Vektor Produkts** ist **parallelisierbar**.

Topic sensitiver Page-Rank

- Die große Schwäche: Der Page-Rank ist anfrage-unabhängig.
- Natürlich arbeitet Google bereits „topic sensitiv“.

Themengebiete des **Open Directory Project**s bieten sich an:

1. *Arts* (Movies, Television, Music,...),
2. *Business* (Jobs, Real Estate, Investing,...),
3. *Computers* (Internet, Software, Hardware,...),
4. *Games* (Video Games, RPGs, Gambling,...),
5. *Health* (Fitness, Medicine,...),
6. *Home* (Family, Consumers, Cooking,...),
7. *Kids and Teens* (School, Teen Life,...),
8. *News* (Media, Newspapers, Weather,...),
9. *Recreation* (Travel, Food, Outdoors, Humor,...),
10. *Reference* (Maps, Education, Libraries,...),
11. *Science* (Computer Science, Biology, Physics,...)

Eine mögliche Personalisierung

- Definiere den Page-Rank für Themengebiet T_i durch

$$\mathrm{pr}^{(i)} = d \cdot p^{(i)} + (1 - d) \cdot \mathrm{pr}^{(i)} \cdot P,$$

  - Bisher haben wir die Gleichverteilung für $p^{(i)}$ gewählt. Jetzt soll $p_w^{(i)}$ die Relevanz von Seite w für Themengebiet T_i wiedergeben.
  - Berechne $p_w^{(i)}$ offline.
- Für eine Suchanfrage $Q = (Q_j \mid j)$:
  - Berechne $\mathrm{prob}[Q_j \mid T_i]$ offline.
  - Ermittle $\mathrm{prob}[T_i \mid Q] = \frac{\prod_j \mathrm{prob}[Q_j \mid T_i] \cdot \mathrm{prob}[T_i]}{\prod_j \mathrm{prob}[Q_j]}$.
- Berechne den „Rang“ der Seite w (über alle ausgewählten Themengebiete) durch

$$\mathrm{Rang}(w \mid Q) = \sum_i \mathrm{prob}[T_i \mid Q] \cdot \mathrm{pr}_w^{(i)}.$$

Hubs und Authorities

Hubs und Authorities

Für eine Suchanfrage σ möchten wir die aussagekräftigsten Webseiten erhalten.

- **Die Idee**: Unterscheide
  - **Hubs**: Seiten mit „guten“ Links
  - und **Authorities**: Aussagekräftige Seiten.

  Bestimme Authorities mit Hilfe der Hubs.
- **Das Problem**:
  Wir kennen anfänglich weder Hubs noch Authorities!
- **Das Vorgehen**: Gehe iterativ vor.
  - Ein Dokument, das auf viele Dokumente mit hohem Authority-Gewicht zeigt, soll ein hohes Hub-Gewicht erhalten,
  - ein Dokument, auf das viele Dokumente mit hohem Hub-Gewicht zeigt, soll ein hohes Authority-Gewicht erhalten.

Hyperlink Induced Topic Search (HITS) (1/3)

Wir treffen zuerst eine Vorauswahl der für σ interessanten Seiten.

1. σ ist die Suchanfrage.
2. Sei W_σ die Menge aller Dokumente, die die Stichworte der Anfrage enthalten.
   - Bestimme eine kleine Teilmenge R_σ relevanter Dokumente mit Hilfe einer textbasierten Suchmaschine.
   - Die besten Seiten sind möglicherweise nicht in R_σ enthalten: Deshalb vergrößern wir die Menge im nächsten Schritt.

3. Setze $S_\sigma = R_\sigma$. Für jedes Dokument $w \in R_\sigma$:
   - Füge alle Dokumente zu S_σ hinzu, auf die w zeigt.

     // Das werden nicht viele Seiten sein:
     // Wir sollten jetzt fast alle Authorities erhalten haben.
   - Füge alle Dokumente zu S_σ hinzu, die auf w zeigen: Wenn zuviele, dann wähle eine beliebige Teilmenge aus.

     // Das oberste Ziel ist die Bestimmung aller Authorities:
     // Wir sorgen dafür, das die gerade gewonnenen Authorities
     // genügend Unterstützung haben.
4. Berechne Hub- und Authority-Gewichte.

Hyperlink Induced Topic Search (HITS) (3/3)

Wir beschränken uns auf den Graphen $G_\sigma = (S_\sigma, E_\sigma)$. Die Kanten entsprechen den Hyperlinks zwischen Dokumenten aus S_σ.

1. Es sei $n = |S_\sigma|$. Setze $A_w = H_w = \frac{1}{\sqrt{n}}$.
   Kommentar: $||A|| = ||H|| = 1$.
2. Wiederhole genügend oft:
   - $H_u = \sum_{w,\ (u,w) \in E_\sigma} A_w$.

     /* Das Hub-Gewicht von u ist groß, wenn u auf viele Dokumente mit hohem Authority-Gewicht zeigt. */
   - $A_u = \sum_{w,\ (w,u) \in E_\sigma} H_w$.

     /* Das Authority-Gewicht von u ist groß, wenn viele Dokumente mit hohem Hub-Gewicht auf u zeigen. */
   - Normalisiere A und H, d.h. setze $A = \frac{A}{||A||}$ und $H = \frac{H}{||H||}$.

Analyse

- Setze $x_0 = (1/\sqrt{n}, \ldots, 1/\sqrt{n})^T$ und
- bezeichne den Authority- bzw. den Hub-Vektor nach der $i$ten Iteration mit A^i bzw. mit H^i.

- Sei M die Adjazenzmatrix von G_σ. Ohne Normalisierung gilt

$$H^{i+2} = M \cdot A^{i+1} \text{ und } A^{i+1} = M^T \cdot H^i.$$

- Und als Konsequenz:

$$H^{i+2} = M \cdot A^{i+1} = M \cdot M^T \cdot H^i \text{ und } A^{i+2} = M^T \cdot H^{i+1} = M^T \cdot M \cdot A^i.$$

- Also ist

$$H^{2k} = (M \cdot M^T)^k \cdot x_0 \text{ und } A^{2k} = (M^T \cdot M)^k \cdot x_0.$$

Beachte: $M \cdot M^T$ und $M^T \cdot M$ sind symmetrische Matrizen!

Symmetrische Matrizen: Was muß man wissen?

Symmetrische Matrizen

Sei K eine symmetrische Matrix mit n Zeilen und Spalten.

(a) Sämtliche Eigenwerte $\lambda_1, \ldots, \lambda_n$ von K sind reellwertig.

(b) Es gibt eine Orthonormalbasis $b_1, \ldots, b_n$ von Eigenvektoren von K. (D.h. es ist $K \cdot b_i = \lambda_i \cdot b_i$ und das innere Produkt $\langle b_i, b_j \rangle$ verschwindet für $i \neq j$ und ist Eins für $i = j$.)

(c) Der betragsmäßig größte Eigenwert λ_1 sei vom Betrag her größer als der betragsmäßig zweitgrößte Eigenwert. Wenn der Vektor x_0 nicht senkrecht auf dem Eigenvektor v von λ steht, dann konvergiert die Folge $(\frac{x_k}{\lambda_1^k} \mid k \in \mathbb{N})$ mit

$$x_{k+1} = \frac{K \cdot x_k}{||K \cdot x_k||}.$$

gegen $\pm v$.

Warum konvergiert die Folge?

- Es gibt eine Orthonormalbasis $v_1, \ldots, v_n$ aus den Eigenvektoren zu den Eigenwerten $\lambda_1 > \lambda_2 \geqslant \cdots \geqslant \lambda_n$.
- Also gibt es eine Linearkombination $x_0 = \sum_{i=1}^{n} \alpha_i \cdot v_i$.
- Weiterhin ist $K \cdot x_0 = \sum_{i=1}^{n} \alpha_i \cdot K \cdot v_i = \sum_{i=1}^{n} \alpha_i \cdot \lambda_i \cdot v_i$ und nach t-facher Iteration

$$K^t \cdot x_0 = \sum_{i=1}^{n} \alpha_i \cdot \lambda_i^t \cdot v_i.$$

Falls $\alpha_1 \neq 0$

Das Gewicht des Eigenvektors v_1 nimmt am stärksten zu, denn

$$\lim_{t \to \infty} \lambda_i^t / \lambda_1^t = 0$$

gilt für $i \neq 1$.

Und die Konsequenzen für das HITS-Verfahren?

In unserer Situation ist $K = MM^T$ oder $K = M^T M$.

- Wenn der größte Eigenwert größer als der zweitgrößte Eigenwert ist und der Vektor x^0 nicht senkrecht auf dem größten Eigenvektor steht, dann konvergieren
  - die HITS Authority-Gewichte bis auf das Vorzeichen gegen den größten Eigenvektor von $M^T M$ und
  - die HITS Hub-Gewichte bis auf das Vorzeichen gegen den größten Eigenvektor von MM^T.
- Übungsaufgabe:
  - Alle Eigenwerte sind nicht-negativ.
  - Die Folge $(x_k \mid k \in \mathbb{N})$ konvergiert immer.
  - Wenn ein Vektor x^0 nur positive Komponenten hat, dann kann x^0 nicht senkrecht auf dem größten Eigenvektor stehen.

HITS versus Page-Rank

- *Berechnungsaufwand:*
  - Der Aufwand für Page-Rank ist gering, da die Bewertungen vorberechnet sind.
  - Für das HITS Verfahren wird experimentell beobachtet, dass 20 Iterationen für Mengen S_σ mit $|S_\sigma| \approx 1000$) „reichen“.
- *Spamming:*
  - Die Page-Rank Bewertung ist global.
  - HITS ist bei einer „verdorbenen“ Auswahl R_σ verloren.
- *Qualität des Suchergebnisses:*
  - Beide Verfahren liefern im Allgemeinen Ergebnisse guter Qualität.
  - HITS bezieht die Suchanfrage in seinem Ranking mitein.

Meta-Suchmaschinen und Social Choice Theorie

Meta-Suchmaschinen

Eine Meta-Suchmaschine gibt eine Anfrage an mehrere Suchmaschinen weiter und berechnet aus den einzelnen Reihenfolgen eine neue Reihenfolge hoffentlich relevanter Dokumente.

Das Integrationsproblem:

- Gegeben sind Teilmengen $T_1, \ldots, T_n \subseteq U$ eines Universums U sowie vollständige Ordnungen (oder Reihenfolgen) $<_1, \ldots, <_n$, wobei $<_i$ eine Reihenfolge auf der Teilmenge T_i ist.
- Es ist eine möglichst „gute“ Reihenfolge für die Teilmenge $T = \bigcup_{i=1}^{n} T_i$ zu bestimmen.

Was sollte man von einer möglichst guten Lösung fordern?

Das Integrationsproblem: Eine Anwendung

Welche Vorschläge sollte ein Flug-Reservierungssystem erstellen, wenn Präferenzen im Hinblick auf

- Preis,
- Reisezeit,
- Reiselänge,
- Anzahl der Zwischenstops,
- Wahl von Fenster- oder Gangsitzen,
- Frequent-Flier Optionen und
- Ticket-Rückgaberecht

bekannt sind?

Social Choice Theorie

Wie sollte man eine kollektive Entscheidung fällen, wenn nur individuelle Präferenzen bekannt sind?

- Wie sollte man Abstimmungen, bzw. Wahlen bei komplexen Sachverhalten organisieren?
  - Was sind komplexe Sachverhalte?
    Nicht eine Option aus zwei Optionen ist zu bestimmen, sondern eine Reihenfolge für mindestens drei Optionen.
- Das kann doch nicht schwer sein!
  Wähle z. B. eine Reihenfolge, die Option **A** höher als Option **B** wertet, wenn eine Mehrheit **A** gegenüber **B** vorzieht.

Das Condorcet Paradox

Drei Personen x, y und z haben verschiedene Vorlieben:

- x bevorzugt die Reihenfolge A, B, C.
- y bevorzugt die Reihenfolge B, C, A, und
- z bevorzugt die Reihenfolge C, A, B.

- Eine Mehrheit bevorzugt Option A vor B, Option B vor Option C aber auch Option C vor Option A.
- Was ist passiert? Mehrheitsentscheidungen sind nicht notwendigerweise transitiv!

Wenn über je zwei Optionen abgestimmt wird, dann erhält man nicht notwendigerweise eine Reihenfolge.

Der Unmöglichkeitssatz von Arrow

Der Unmöglichkeitssatz von Arrow (1/2)

Wir betrachten Reihenfolgen auf einem Universum U.

Ein **Präferenzen-Funktional** P ordnet n Reihenfolgen $<_1, \ldots, <_n$ eine Reihenfolge $<$ zu.

- P ***respektiert Einstimmigkeit***, falls für alle $x, y \in U$ gilt

$$x <_i y \text{ für alle } i \in \{1, \ldots, n\} \quad \Rightarrow \quad x < y.$$

- P ist ***unabhängig von irrelevanten Alternativen***, wenn für alle $x, y \in U$ nur die n individuellen Vergleiche zwischen x und y über die Präferenz $x<y$ entscheiden.

- Gibt es Funktionale, die Einstimmigkeit respektieren und unabhängig von irrelevanten Attributen sind?
- Ja, das ***Diktatur-Funktional*** $D_i(<_1, \ldots, <_n) = <_i$.

Der Unmöglichkeitssatz von Arrow (2/2)

Es gelte $|\mathbf{U}| \geqslant \mathbf{3}$. Wenn ein Präferenzen-Funktional P unabhängig von irrelevanten Alternativen ist und Einstimmigkeit respektiert, dann ist P ein **Diktatur-Funktional**.

Und nun?

- Unser Funktional sollte Einstimmigkeit respektieren, oder?
- Für die Entscheidung zwischen zwei Optionen sollte doch eine dritte Option keine Rolle spielen, oder?!

Unabhängigkeit von irrelevanten Attributen?!

- Als Nachtisch wurden dem amerikanischem Philosophen Sidney Morgenbesser Apfel- und Blaubeerkuchen angeboten.
  - Er entschied sich für den Apfelkuchen.
  - Als die Kellnerin ergänzte, das auch Erdbeerkuchen im Angebot sei, nahm er den Blaubeerkuchen.
- Aber jetzt im Ernst: Und wenn zwei Optionen starke Rangunterschiede in den einzelen Reihenfolgen aufweisen?
  - Diese Rangunterschiede entstehen durch „Dritte“ und dürfen deshalb nicht berücksichtigt werden!?

Der Satz von Gibbard-Satterthwaite (1/2)

Gibt es zumindest eine überzeugende „**soziale Option**“?

Das Funktional Q weise einer Folge $<_1, \ldots, <_n$ von Ordnungen eine Option $u \in U$ zu. Das Funktional Q heißt **monoton**, wenn gilt

$$Q(<_1, \ldots, <_i, \ldots, <_n) = a \neq b = Q(<_1, \ldots, \prec_i, \ldots, <_n)$$
$$\Rightarrow \quad b <_i a \ \wedge \ a \prec_i b.$$

Wenn sich die gewählte Option ändert, weil eine Reihenfolge geändert wurde, dann

- muss die neue Reihenfolge die neue Option gegenüber der alten
- und die alte Reihenfolge die alte Option gegenüber der neuen vorziehen.

Ist Q nicht monoton, dann kann es vorteilhaft sein, sich taktisch zu verhalten und gegen die eigenen Überzeugungen zu stimmen: Warum für eine chancenlose Option stimmen?

Der Satz von Gibbard-Satterthwaite (2/2)

- Es gelte $|\mathbf{U}| \geqslant \mathbf{3}$.
- Q sei ein monotones Funktional, das jede Option in U mindestens einmal als Wert annehmen kann.

Dann ist Q ein **Diktator-Funktional**:
Q wählt stets die beliebteste Option einer **fixierten** Eingabe-Ordnung.

Es ist zu entscheiden, welches von mindestens drei Baumaßnahmen mit Steuergeldern zu finanzieren ist.

- Für die Entscheidung ist die ehrliche Angabe der Präferenzen eine Vorbedingung.
- Nach dem Satz von Gibbard-Satterthwaite gibt es aber kein Entscheidungsverfahren, das taktisches, unehrliches Verhalten ausschließen kann.

Gibbard-Satterthwaite als Konsequenz von Arrow

Angenommen Q ist ein monotones Funktional mit Eingaben $<_1, \ldots, <_n$.

- Wir erzeugen ein Präferenzen-Funktion P aus Q und zeigen, dass P Einstimmigkeit respektiert und unabhängig von irrelevanten Attributen ist.

  Die Behauptung folgt dann aus dem Satz von Arrow.
- Um die Ordnung $P(<_1, \ldots, <_n) = <$ festzulegen, genügt es für je zwei Optionen $u, v \in U$ zu klären, ob $u < v$ oder $v < u$ gilt.
  - Dazu bewegen wir u und v in jeder Ordnung $<_i$ ganz nach oben, behalten die ursprüngliche Präferenz zwischen u und v aber bei.
  - Wenn $\prec_1, \ldots, \prec_n$ die neuen Ordnungen sind, dann setze

$$u < v \;\Leftrightarrow\; Q(\prec_1, \ldots, \prec_n) = v.$$

  - **Zeige**, dass P Einstimmigkeit respektiert und unabhängig von irrelevanten Attributen ist.

Borda's Regel, die Kendall- und die Spearman Distanz

Borda's Regel

- Für jedes $x \in U$ ist

$$\text{Rang}_i(x) = |\{z \in U \mid z \leqslant_i x\}|$$

der Rang von x bezüglich $<_i$.

- Bestimme eine Reihenfolge gemäß steigender Rangsumme $\sum_{i=1}^{n} \text{Rang}_i(x)$.

- Borda's Regel respektiert **Einstimmigkeit**, verletzt aber **Unabhängigkeit von irrelevanten Alternativen**:

  Rangunterschiede sind in Borda's Regel entscheidend und werden durch „Dritte" hervorgerufen.

- Borda's Regel verletzt auch die **abgeschwächte Demokratie-Eigenschaft** (ADE):

  Wenn für irgendeine Teilmenge $S \subseteq U$ und für alle $y \in S$ und $x \in \overline{S}$ stets eine Mehrheit der Reihenfolgen y gegenüber x bevorzugt, dann ist $x < y$.

Die Kendall Distanz

ADE: Wenn für eine Teilmenge $S \subseteq U$ und für alle $y \in S, x \in \overline{S}$ stets eine Mehrheit der Reihenfolgen y über x bevorzugt, dann ist $x < y$.

- Wenn ADE gilt, kann Spamming unterdrückt werden, solange eine Mehrheit der Suchmaschinen Spamming-Versuche erkennt.
- Kendall-Reihenfolgen erfüllen ADE (Beweis später):
  - Die **Kendall-Distanz** der Reihenfolgen $<_1$ und $<_2$ ist

$$K(<_1, <_2) = |\{(u, v) \mid u <_1 v, v <_2 u\}|.$$

  - Eine **Kendall-Reihenfolge** $<$ für $<_1, \ldots, <_n$ minimiert die Summe

$$\sum_{i=1}^{n} K(<_i, <)$$

    der Kendall-Distanzen.

Kendall-Reihenfolgen erfüllen ADE

- Leider führt die Bestimmung einer Kendall-Reihenfolge auf ein $\mathcal{NP}$-vollständiges Problem.
- Aber jede Approximation $<$ der Kendall-Reihenfolge kann weiter verbessert werden, so dass ADE gilt.

Die Reihenfolgen $<_1, \ldots, <_n$ und $<$ seien vorgegeben. Dann kann eine Reihenfolge $<^*$ in Zeit $O(n \cdot |\bigcup_{i=1}^{n} T_i|^2)$ bestimmt werden, so dass

- die Reihenfolge $<^*$ ADE erfüllt und
- $\sum_i K(<_i, <^*) \leqslant \sum_i K(<_i, <)$ gilt.

Die Verbesserungsstrategie

Die Reihenfolgen $<_1, \ldots, <_n$ und $<$ seien vorgegeben.

- O.B.d.A. gelte $U = \{1, \ldots, |U|\}$ und $1 > 2 > \cdots > |U|$.
- Die neue Reihenfolge $<^*$ sei bereits auf $\{1, \ldots, k\}$ definiert.
  - Vertausche $k+1$ mit dem bzgl. $<^*$ k-kleinsten Element, wenn eine Mehrheit der Reihenfolgen dies verlangt.

    Die Kendall-Distanz kann höchstens fallen.
  - Wiederhole, falls notwendig.

Die Reihenfolge $<^*$ erfüllt ADE. Warum?

- Wenn nicht, dann ist eine Menge $S \subseteq U$ ein Gegenbeispiel:

  Eine Mehrheit bevorzugt $s \in S$ über $t \in \overline{S}$, aber „$S > \overline{S}$“ gilt nicht.
- $u \in S$, $v \notin S$ sei ein Gegenbeispiel mit geringster Distanz bzgl. $<^*$. Insbesondere ist also $u <^* \cdots <^* w <^* v$.
  - Es ist $w \in S$: Sonst ist u, w ein Gegenbeispiel kleinerer Distanz.
  - Also bevorzugt eine Mehrheit w über v: Unsere Verbessserungs-strategie würde aber w und v vertauschen: **Widerspruch**.

Eine Approximation der Kendall-Reihenfolge

- Die **Spearman-Distanz** $S(<_1, <_2) = \sum_{x \in U} |\text{Rang}_1(x) - \text{Rang}_2(x)|$ summiert Rangunterschiede auf.
- Eine **Spearman-Reihenfolge** $<$ minimiert $\sum_{i=1}^{n} S(<_i, <)$.

- Übungsaufgabe: $K(<_1, <_2) \leqslant S(<_1, <_2) \leqslant 2 \cdot K(<_1, <_2)$. Also ist eine Spearman-Reihenfolge 2-approximativ.
- Eine Spearman-Reihenfolge kann effizient bestimmt werden.
  - Betrachte den vollständigen bipartiten Graphen mit Knotenmengen $V_1 = U$ und $V_2 = \{1, \ldots, |U|\}$.

    Für ein Element $x \in V_1$ und eine „Position“ $p \in V_2$ füge die Kante $\{x, p\}$ mit Gewicht $\sum_{i=1}^{n} |\text{Rang}_i(x) - p|$ ein.
  - Ordnungen $R \Leftrightarrow$ perfekte Matchings M_R:
    - ★ Das Kantengewicht des Matchings M_R stimmt mit der Spearman-Distanz von R überein.
  - Ein perfektes Matching mit minimalem Gewicht ist effizient konstruierbar.

Zusammenfassung

- **Connectivity Analysis**:
  - Google bestimmt die stationäre Verteilung der Web-Kette und benutzt sie als Page-Rank.
  - Der HITS-Algorithmus bestimmt Authorities (Seiten hoher Qualität für eine Suchanfrage) mit Hilfe von Hubs (Seiten mit guten Links).
- Das **Integrationsproblem**:
  - Der Satz von Arrow zeigt, dass überzeugende Lösungen nicht existieren.
  - Kendall-Reihenfolgen erfüllen zumindest die abgeschwächte Demokratie-Eigenschaft.
  - Die schwierig zu bestimmenden Kendall-Reihenfolgen können durch Spearman-Reihenfolgen approximiert werden.
  - Sind Kendall-Reihenfolgen überzeugende Lösungen?