Max-Planck-Institut Saarbrücken
AG5: Datenbanken und Informationssysteme
Dozenten: Prof. Dr.-Ing. G. Weikum, Dr. Ralf Schenkel
Betreuer: Dipl.-Inform. Sergej Sizov, Dipl.-Inform. Christian Zimmer

Informationssysteme (SS 04)
Übungsblatt 3

Ausgabe: 11.Mai 2004

Abgabe: 18.Mai 2004 in der Vorlesung

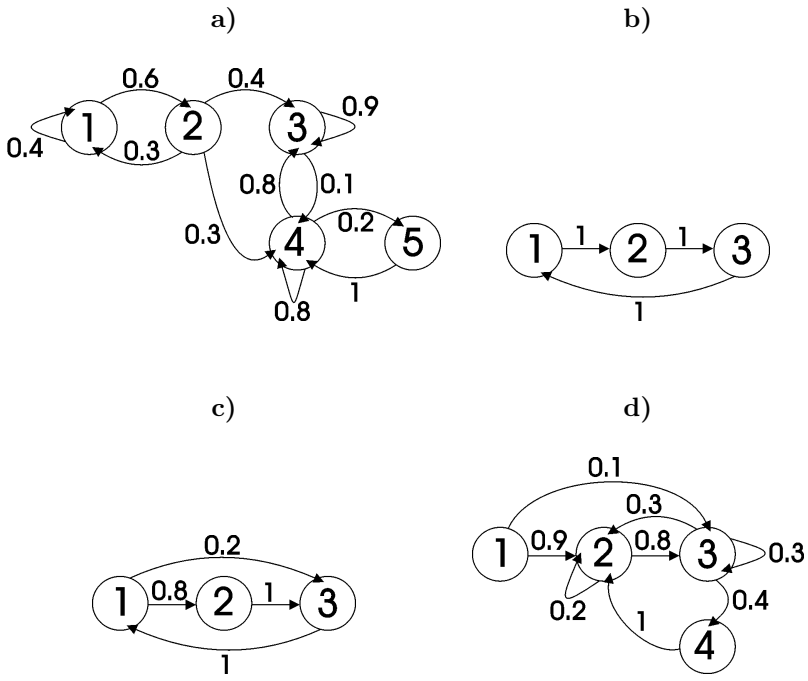
Aufgabe 3.1: Markov-Ketten

a) Konstruieren Sie ein Beispiel für eine periodische Markov-Kette, welche nicht ergodisch ist und zeigen Sie, dass die Zustandswahrscheinlichkeiten nach n-Schritten nicht zu konstanten Wahrscheinlichkeiten konvergieren.

b) Konstruieren Sie ein Beispiel einer reduzierbaren Markov-Kette, welche nicht ergodisch ist und zeigen Sie, dass die Zustandswahrscheinlichkeiten nicht zu konstanten Wahrscheinlichkeiten konvergieren, die von der anfänglichen Verteilung unabhängig ist.

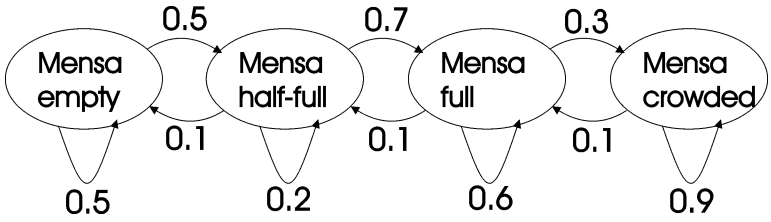
Aufgabe 3.2:

Welche der folgenden Markov-ketten sind irreducible, aperiodic oder ergodic?



Aufgabe 3.3: Markov-Ketten

Betrachten Sie die folgende Markov-Ketten mit 4 Zuständen:



a) Zeigen Sie, dass die Kette ergodisch ist.

b) Berechnen Sie die stationären Zustandswahrscheinlichkeiten.

Aufgabe 3.4: Anwendung von Markov-Modellen

Betrachten Sie einen Web Server, der Webseiten auf der Festplatte speichert und häufig genutzte Seiten im Hauptspeicher behält. Nehmen Sie an, dass alle Seiten gleich groß sind und es insgesamt N Seiten sind; die Cachsgroße ist $M < N$. Nehmen Sie ferner an, dass Seiten mit festen Wahrscheinlichkeiten durch Benutzer aufgerufen werden, welche unabhängig sind von der Seite selbst und den Seitenzugriffen. Nehmen Sie auch an, dass die Zugriffswahrscheinlichkeiten einer *Zipf* Verteilung folgen und dass die Seiten in absteigender Reihenfolge ihrer Zugriffswahrscheinlichkeit nummeriert sind:

$$P[n\ddot{a}chster\ Zugriff\ verlangt\ Seite\ i] \quad = \quad \frac{(1/i)}{\sum_{i=1}^{N} 1/i}$$

Betrachten Sie die Situation, dass der Cache gefüllt ist und eine Seite wird angefragt, die sich nicht im Cache befindet. In diesem Fall wird vom Cache-Manager eine Seite aus dem Cache enfernt, indem eine beliebige Seite ausgewählt wird - folgend einer Gleichverteilung. Diese Ersetzungsstrategie wird als *random policy* bezeichnet.

Nun betrachten Sie das Problem, die Trefferquote des Caches vorherzusagen, d.h. die Wahrscheinlichkeit, dass die nächste Anfrage eine Seite aus dem Cache betrifft. Zur Vereinfachung gehen Sie vom speziellen Fall mit $N = 4$ und $M = 2$ aus.

a) Modellieren Sie die möglichen Zustände des Caches und deren übergänge als eine Markov-Kette.

b) Berechnen Sie die stationären Zustandswahrscheinlichkeiten und benutzen Sie diese, um die Trefferquote des Cahces vorherzusagen.

c) Nehmen sie ferner an, dass die Seiten im Cache eine time-to-live Zeit T zugeordnet bekommen, nach der die Seite auf jeden Fall aus dem Cache entfernt wird. Dies soll dazu dienen, dass der Cache aktuell ist - niemals älter als T Zeitschritte. Betrachten Sie nun wieder die Fragestellungen a) und b) für dieses erweiterte Model.

Anmerkung: Dieser Teil der Aufgabe kann etwas schwieriger sein.

Aufgabe 3.5: Page-Rank-Verfahren

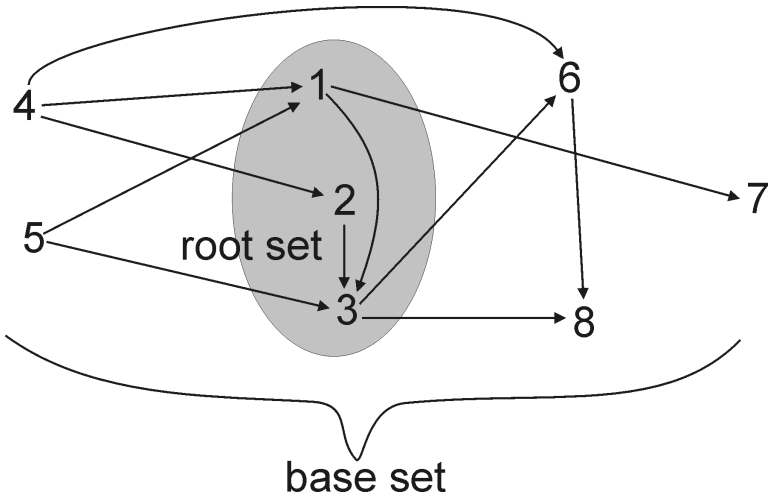
Betrachten Sie den Graphen $G = (V, E)$ mit der Knotenmenge $V = \{1, 2, 3, 4\}$ und der Kantenmenge $E = \{(1, 2), (2, 3), (3, 4), (4, 2)\}$. Bestimmen Sie die Überführungsmatrix P für den *Random Walk* eines Web-Surfers with $\epsilon = 0.01$. Berechnen Sie den Page Rank von V

1) iterativ mit initialen Werten $r(1) = r(2) = r(3) = r(4) = 0.25$ und 4 Iterationsstufen

2) direkt durch Berechnung des Eigenvektors, d.h. indem ein Lineares Gleichungssystem geöst wird

Aufgabe 3.6: HITS

Wenden Sie den HITS Algorithmus (in seiner einfachsten Form ohne Kantengewichte usw.) auf die folgende Grundmenge an:



Aufgabe 3.7: Erweitertes HITS

Betrachten Sie den HITS Algorithmus mit *mutual reinforcement* zwischen guten Hubs und guten Autoritäten. Nehmen Sie an, dass viele gute Hubs heterogen im folgenden Sinne sind: sie enthalten mehrere Sammlungen von Links auf verschiedene Themenbereiche, wobei die meisten Links auf gute Autoritäten verweisen während ansdere bestenfalls mittelmäßig sind. Ein Beispiel hierzu könnte die Homepage eines Forschers sein mit guten Links über seine Forschungsthemen und mittelmäßigen Links zu seinen Hobbies.

a) Diskutieren Sie, zu welchen weiteren Problemen dies im HITS Algorithmus führen kann. Überlegen Sie sich ein Szenario, das zu Annomalien führt oder zeigen Sie, dass der Algorithmus robust genug ist, um mit solchen Fällen umzugehen - z.B. durch ein kanonisches Beispiel.

b) Überlegen sie sich Ansätze für diese heterogenen Hubs. Sie können annehmen, dass alle Seiten in HTML vorliegen und dass die Verknüpfungen zu verschiedenen Themen durch die HTML-Tag-Struktur unterteilt werden kann - z.B. durch Überschriften wie *Forschungsthemen* oder *Hobbies*.

Aufgabe 3.8: HITS mit SVD

Zeigen Sie, dass die HITS Berechnung von Authority-Scores auch durch Anwendung von SVD auf eine geeignete Matrix erreicht werden kann.

Aufgabe 3.9: Sonderfälle bei Page-Rank und HITS

a) Welchen Einfluss haben Knoten mit Eingangsgrad 0 beim Page-Rank- und beim HITS-Verfahren? Welchen Einfluss haben Knoten mit Ausgangsgrad 0? Ergeben sich daraus ggf. Optimierungen fr die Berechnung der Autoritäts-Scores?

b) Lassen sich das Page-Rank-Verfahren und das HITS-Verfahren auch auf ungerichtete Graphen anwenden? Macht das Sinn?

Aufgabe 3.10: Themenspezifisches HITS-Verfahren

Entwerfen Sie - analog zum themenspezifischen Page-Rank-Verfahren - eine Methode, die themenspezifische Authorities und Hubs nach einem geeignet erweiterten HITS-Verfahren ermittelt.