Google Pagerank

Hvordan man finder en nål i en høstak

Georg Mohr, 4. marts 2008

Kim Knudsen

kim@math.aau.dk

Institut for Matematiske Fag

Aalborg Universitet

http://www.math.aau.dk/~kim/GeorgMohr2008.pdf

Internettet - informationsalderen

- Moderne teknologis største vidunder: Kommunikation, globalisering, information, demokrati?...

Internettet - informationsalderen

- Moderne teknologis største vidunder: Kommunikation, globalisering, information, demokrati?...
- Historie:

  1957: USSR sender Sputnik i kredsløb om jorden → US DoD etablerer Advanced Research Projects Agency (’58)

  1969: ARPANET. Første netværk med *packet switching* (modsætning til *circuit switching*).

  1988: Internettet åbnes for kommercielle interesser

  1991: World Wide Web lanceres

  2008: www har 1.1 mia. brugere og 20 mia. hjemmesider

Internettet - informationsalderen

- Moderne teknologis største vidunder: Kommunikation, globalisering, information, demokrati?...
- Historie:

  1957: USSR sender Sputnik i kredsløb om jorden → US DoD etablerer Advanced Research Projects Agency (’58)

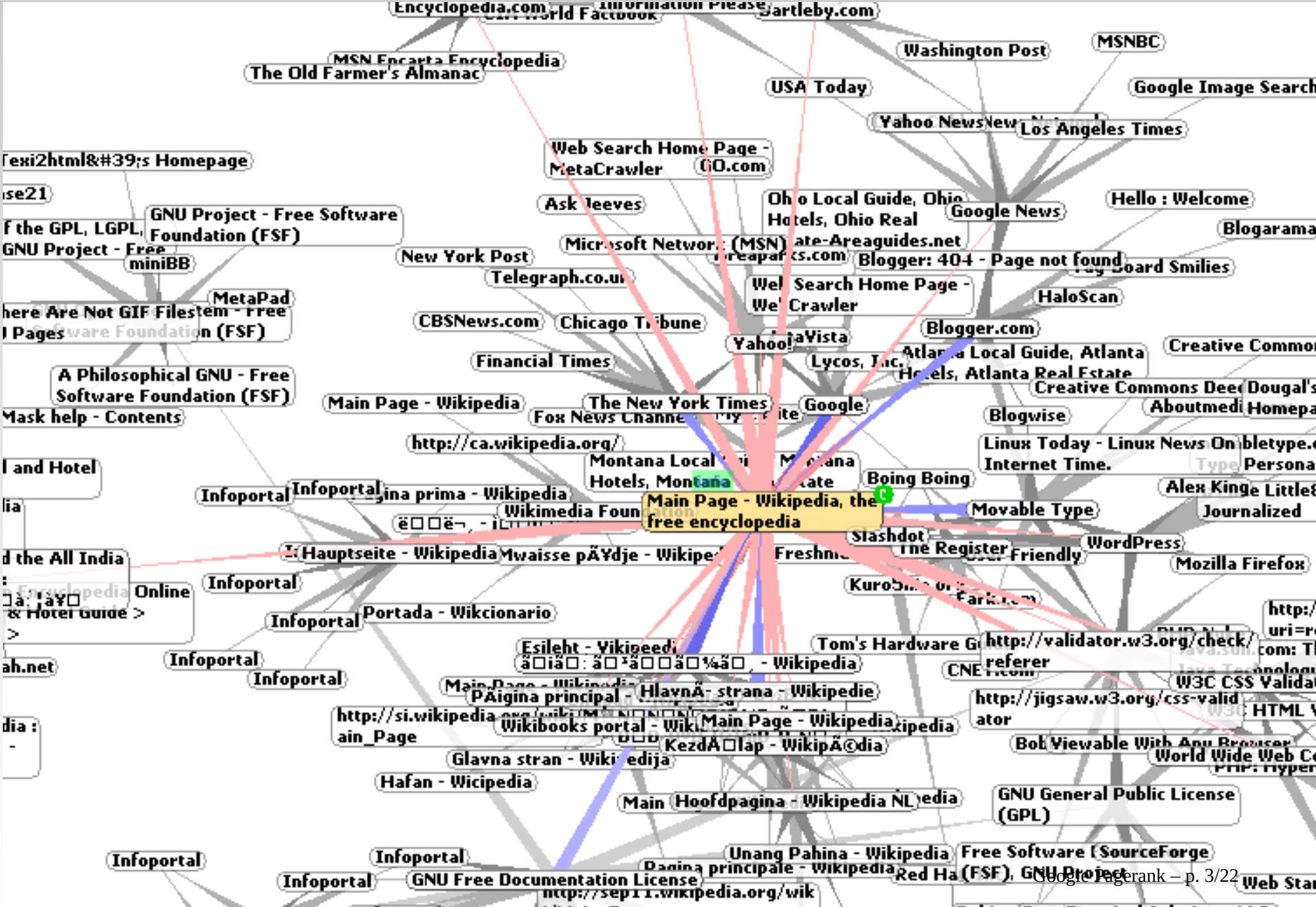
  1969: ARPANET. Første netværk med *packet switching* (modsætning til *circuit switching*).

  1988: Internettet åbnes for kommercielle interesser

  1991: World Wide Web lanceres

  2008: www har 1.1 mia. brugere og 20 mia. hjemmesider
- Internettet indeholder alverdens viden og information(?).
- Informationsøgning: Hvordan finder man en nål i en høstak?
  Søgemaskiner: yahoo!, msn, google...

Udsnit af internettet (0.0001%)



Google Inc., kort fortalt

- Google er (et af) verdens mest succesfyldte IT-firmaer med hovedkontor i Mountain View, CA. Grundlagt i 1998 af Larry Page og Sergey Brin

Google Inc., kort fortalt

- Google er (et af) verdens mest succesfyldte IT-firmaer med hovedkontor i Mountain View, CA. Grundlagt i 1998 af Larry Page og Sergey Brin

- Mission: “At organisere verdens information og gøre den universelt tilgængelig”
  Motto: “Don’t do evil”!

Google Inc., kort fortalt

- Google er (et af) verdens mest succesfyldte IT-firmaer med hovedkontor i Mountain View, CA. Grundlagt i 1998 af Larry Page og Sergey Brin

- Mission: “At organisere verdens information og gøre den universelt tilgængelig”
  Motto: “Don’t do evil”!
- Verdens mest brugte søgemaskine, google.com (>50% af alle søgninger)

Google Inc., kort fortalt

- Google er (et af) verdens mest succesfyldte IT-firmaer med hovedkontor i Mountain View, CA. Grundlagt i 1998 af Larry Page og Sergey Brin

- Mission: “At organisere verdens information og gøre den universelt tilgængelig”
  Motto: “Don’t do evil”!
- Verdens mest brugte søgemaskine, google.com (>50% af alle søgninger)
- Andre web-baserede applikationer: Google earth, gmail, google groups, google video, google Picasa, Froogle...

Google Inc., kort fortalt

- Google er (et af) verdens mest succesfyldte IT-firmaer med hovedkontor i Mountain View, CA. Grundlagt i 1998 af Larry Page og Sergey Brin

- Mission: “At organisere verdens information og gøre den universelt tilgængelig”
  Motto: “Don’t do evil”!
- Verdens mest brugte søgemaskine, google.com (>50% af alle søgninger)
- Andre web-baserede applikationer: Google earth, gmail, google groups, google video, google Picasa, Froogle...
- Pengene tjenes idag primært på reklamer ifm. søgemaskinen, gmail mm. I fremtiden?

Google Inc., kort fortalt

- Google er (et af) verdens mest succesfyldte IT-firmaer med hovedkontor i Mountain View, CA. Grundlagt i 1998 af Larry Page og Sergey Brin

- Mission: “At organisere verdens information og gøre den universelt tilgængelig” Motto: “Don’t do evil”!
- Verdens mest brugte søgemaskine, google.com (>50% af alle søgninger)
- Andre web-baserede applikationer: Google earth, gmail, google groups, google video, google Picasa, Froogle...
- Pengene tjenes idag primært på reklamer ifm. søgemaskinen, gmail mm. I fremtiden?
- Hvorfor navnet google? Googol $= 10^{100}$

Googles succes

- Vi er vilde med googles søgemaskine fordi
  1. Den er simpel
  2. Den er hurtig
  3. Den kommer med de “rigtige” svar på selv håbløse forespørgsler

Googles succes

- Vi er vilde med googles søgemaskine fordi
  1. Den er simpel
  2. Den er hurtig
  3. Den kommer med de “rigtige” svar på selv håbløse forespørgsler
- Er der penge i det?
  - Larry Page og Sergey Brin er begge nr. 26 på Forbes liste over verdens rigeste. Er begge gode for 16.6 mia. USD.

Googles succes

- Vi er vilde med googles søgemaskine fordi
  1. Den er simpel
  2. Den er hurtig
  3. Den kommer med de “rigtige” svar på selv håbløse forespørgsler
- Er der penge i det?
  - Larry Page og Sergey Brin er begge nr. 26 på Forbes liste over verdens rigeste. Er begge gode for 16.6 mia. USD.
  - Google inc. tjente i 4. kvartal 2007 ca. 6 mia. $ (skuffende resultat...)

Googles succes

- Vi er vilde med googles søgemaskine fordi
  1. Den er simpel
  2. Den er hurtig
  3. Den kommer med de “rigtige” svar på selv håbløse forespørgsler
- Er der penge i det?
  - Larry Page og Sergey Brin er begge nr. 26 på Forbes liste over verdens rigeste. Er begge gode for 16.6 mia. USD.
  - Google inc. tjente i 4. kvartal 2007 ca. 6 mia. $ (skuffende resultat...)
  - Microsoft bød fornylig 225 mia. kr. for yahoo

Princippet i googles søgemaskine

1. WWW kravles igennem af en web-crawler (web-spider). For enhver side som nås gemmes bl.a. sidens adresse, sidens tekst og sidens links (referencer til andre hjemmesider). Data gemmes i gigantisk database (antal sider $n \sim 20 \cdot 10^9$)
2. Enhver side tildeles et tal (1-10), sidens PageRank, som angiver sidens vigtighed
3. Ved en forespørgsel på google.com findes i databasen de sider, som indeholder søgeordene, og siderne præsenters for forespørgeren i en rækkefølge afhængig af (blandt andet) PageRank

Princippet i googles søgemaskine

1. WWW kravles igennem af en web-crawler (web-spider). For enhver side som nås gemmes bl.a. sidens adresse, sidens tekst og sidens links (referencer til andre hjemmesider). Data gemmes i gigantisk database (antal sider $n \sim 20 \cdot 10^9$)
2. Enhver side tildeles et tal (1-10), sidens PageRank, som angiver sidens vigtighed
3. Ved en forespørgsel på google.com findes i databasen de sider, som indeholder søgeordene, og siderne præsenters for forespørgeren i en rækkefølge afhængig af (blandt andet) PageRank

Nøglen bag succesen er Googles (patenterede) PageRank algoritme. En algoritme der baserer sig på matematik

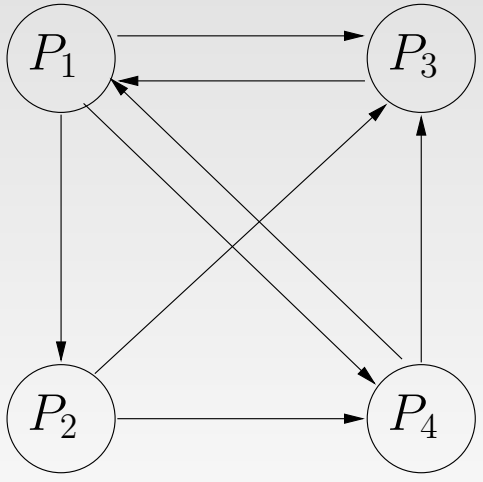
Matematisk model for WWW

WWW modelleres ved en orienteret (multi-)graf: Enhver hjemmeside repræsenteres ved et punkt i grafen.
Hvis der er et link fra siden P_i til siden P_j så repræsenteres det ved en orienteret kant fra punktet hørende til P_i til punktet hørende til P_j.

Matematisk model for WWW

WWW modelleres ved en orienteret (multi-)graf: Enhver hjemmeside repræsenteres ved et punkt i grafen.
Hvis der er et link fra siden P_i til siden P_j så repræsenteres det ved en orienteret kant fra punktet hørende til P_i til punktet hørende til P_j.

Eksempel på mini-WWW:



Hvilken side i netværket er den vigtigste?

Tre basale principper

PageRank bygger på tre demokratiske principper:

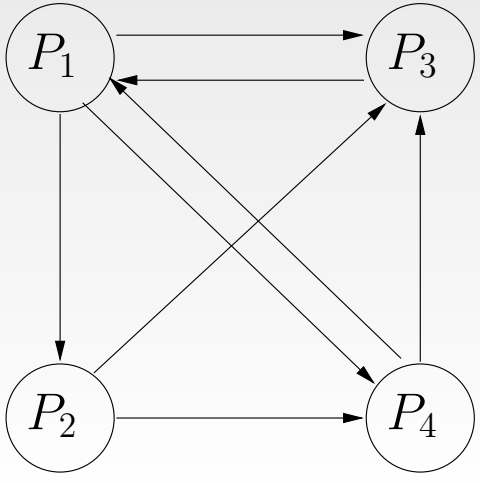
1. Ethvert link tæller som en stemme. En sides vigtighed afhænger af, hvormange andre sider, der stemmer på den
2. Jo større vigtighed en side har, jo større vægt tildeles et udgående link fra siden
3. Jo færre links (=stemmer), der går ud fra en side, jo større vægt tildeles hvert link

Tre basale principper

PageRank bygger på tre demokratiske principper:

1. Ethvert link tæller som en stemme. En sides vigtighed afhænger af, hvormange andre sider, der stemmer på den
2. Jo større vigtighed en side har, jo større vægt tildeles et udgående link fra siden
3. Jo færre links (=stemmer), der går ud fra en side, jo større vægt tildeles hvert link

Hvis x_i angiver vigtigheden af side P_i :



$$x_1 = x_3/1 + x_4/2$$

$$x_2 = x_1/3$$

$$x_3 = x_1/3 + x_2/2 + x_4/2$$

$$x_4 = x_1/3 + x_2/2$$

Lineære ligningssystemer

Reduktion

$$
\begin{aligned}
x_1 &= x_3 + \frac{x_4}{2} && \Leftrightarrow && 0 = -x_1 + x_3 + \frac{x_4}{2} \\
x_2 &= \frac{x_1}{3} && \Leftrightarrow && 0 = \frac{x_1}{3} - x_2 \\
x_3 &= \frac{x_1}{3} + \frac{x_2}{2} + \frac{x_4}{2} && \Leftrightarrow && 0 = \frac{x_1}{3} + \frac{x_2}{2} - x_3 + \frac{x_4}{2} \\
x_4 &= \frac{x_1}{3} + \frac{x_2}{2} && \Leftrightarrow && 0 = \frac{x_1}{3} + \frac{x_2}{2} - x_4
\end{aligned}
$$

Det er et kvadratisk, homogent lineært ligningssystem.

Sådanne ligningssystemer kan have enten én (i givet fald $(0, 0, 0, 0)$) eller uendeligt mange løsninger.

PageRank løsningen

- Vores ligningssystem har uendeligt mange løsninger på formen

$$(x_1, x_2, x_3, x_4) = t(12, 4, 9, 6), \qquad t \in .$$

- Normalisering: Vi vælger den positive løsning, som giver samlet PageRank en:

$$x_1 + x_2 + x_3 + x_4 = 1 \Leftrightarrow (x_1, x_2, x_3, x_4) = \frac{1}{31}(12, 4, 9, 6)$$

- PageRank løsningen er altså

$$(x_1, x_2, x_3, x_4) = \frac{1}{31}(12, 4, 9, 6)$$

Lineær algebra

Hvis vi introducerer link-matricen

$$A = \begin{bmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

så er det opskrevne ligningssystem ækvivalent med

$$Ax = x, \qquad x = [x_1, x_2, x_3, x_4]'.$$

Lineær algebra

Hvis vi introducerer link-matricen

$$A = \begin{bmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

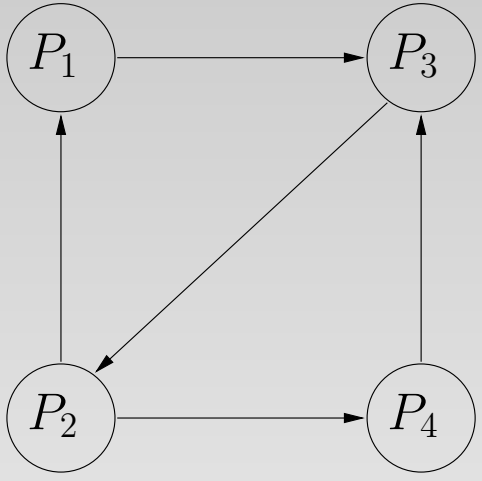
så er det opskrevne ligningssystem ækvivalent med

$$Ax = x, \qquad x = [x_1, x_2, x_3, x_4]'.$$

Den beskrevne problemstilling kan behandles med redskaber fra den del af matematikken, der hedder lineær algebra.
Lineær algebra bruges overalt! Mobiltelefon, ipod, computerspil, beregning af vejrudsigt, MP3-afspiller, digitalkamera...

Opgave

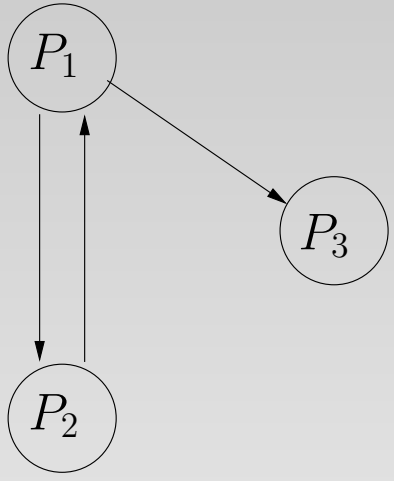
Betragt netværket



- Opskriv de fire ligninger med fire ubekendte (x_1, x_2, x_3, x_4), som kan bruges til at bestemme PageRank for netværket.
- Løs de fire ligninger og find den normaliserede PageRank løsning

Problem 1: Hængende sider

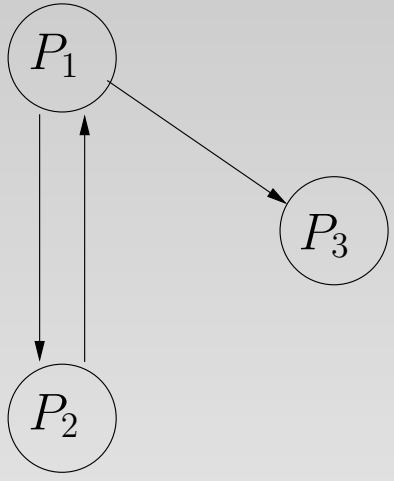
Netværket



giver anledning til ligningerne

Problem 1: Hængende sider

Netværket



giver anledning til ligningerne

$$x_1 = x_2$$
$$x_2 = \frac{x_1}{2}$$
$$x_3 = \frac{x_1}{2}$$

Kan vi finde en normaliseret løsning?

Ny stokstisk model: En tilfældig surfer

- En www-surfer bevæger sig tilfældigt rundt på nettet
- Fra en side P med N udgående links vil han med lige fordelt sandsynlighed $(1/N)$ vælge mellem de udgående links
- PageRank løsningen $x = (x_1, x_2, x_3, \ldots, x_n)$ kan fortolkes som en sandsynlighedsfordeling, hvor x_i angiver sandsynligheden for, at surferen til et vilkårligt tidspunkt (efter lang tid) befinder sig på siden P_i.

Ny stokstisk model: En tilfældig surfer

- En www-surfer bevæger sig tilfældigt rundt på nettet
- Fra en side P med N udgående links vil han med lige fordelt sandsynlighed $(1/N)$ vælge mellem de udgående links
- PageRank løsningen $x = (x_1, x_2, x_3, \ldots, x_n)$ kan fortolkes som en sandsynlighedsfordeling, hvor x_i angiver sandsynligheden for, at surferen til et vilkårligt tidspunkt (efter lang tid) befinder sig på siden P_i.

Problem: Hængende sider, sider som ikke indeholder links.

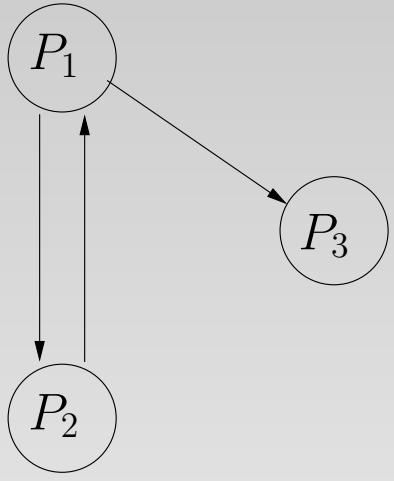
Ny stokstisk model: En tilfældig surfer

- En www-surfer bevæger sig tilfældigt rundt på nettet
- Fra en side P med N udgående links vil han med lige fordelt sandsynlighed $(1/N)$ vælge mellem de udgående links
- PageRank løsningen $x = (x_1, x_2, x_3, \ldots, x_n)$ kan fortolkes som en sandsynlighedsfordeling, hvor x_i angiver sandsynligheden for, at surferen til et vilkårligt tidspunkt (efter lang tid) befinder sig på siden P_i.

Problem: Hængende sider, sider som ikke indeholder links.
Løsning (googles): Fra enhver side uden udgående links indsættes virtuelle links til enhver anden side i netværket.

Løsning: Hængende sider

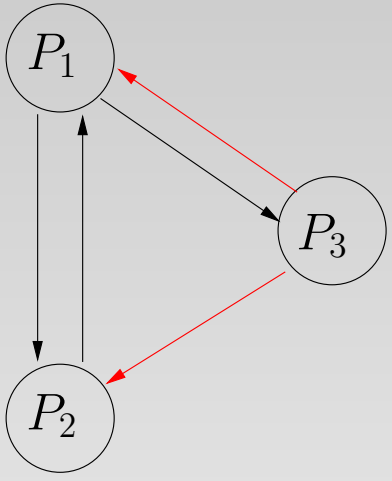
For netværket



indsætter vi virtuelle links fra P_3 til P_1 og P_2.

Løsning: Hængende sider

For netværket



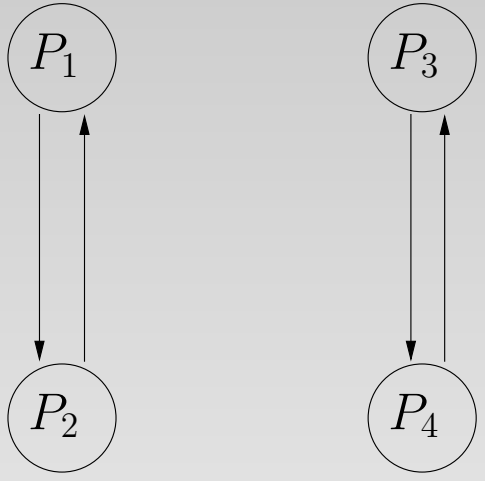
indsætter vi virtuelle links fra P_3 til P_1 og P_2. Det giver os

$$x_1 = x_2 + \frac{x_3}{2}$$
$$x_2 = \frac{x_1}{2} + \frac{x_3}{2}$$
$$x_3 = \frac{x_1}{2}$$

Den normaliserede løsning $(x_1, x_2, x_3) = \frac{1}{9}(4, 3, 2)$

Problem 2: Usammenhængende netværk

Betragt netværket



- Hvilke ligninger giver netværket anledning til?
- Findes der en entydig normaliseret løsning?

Løsning: Usammenhængende netværk

En yderligere modifikation: Når vores www-surfer står på side P, så giver vi ham to muligheder:

1. Med sandsynlighed α vil han som tidligere følge et udgående link fra P
2. Med sandsynlighed $(1 - \alpha)$ vil han springe til en vilkårlig side i netværket

Løsning: Usammenhængende netværk

En yderligere modifikation: Når vores www-surfer står på side P, så giver vi ham to muligheder:

1. Med sandsynlighed α vil han som tidligere følge et udgående link fra P
2. Med sandsynlighed $(1-\alpha)$ vil han springe til en vilkårlig side i netværket

Det svarer til at der i ligningen for ethvert x_j på højresiden ganges med α og leddet

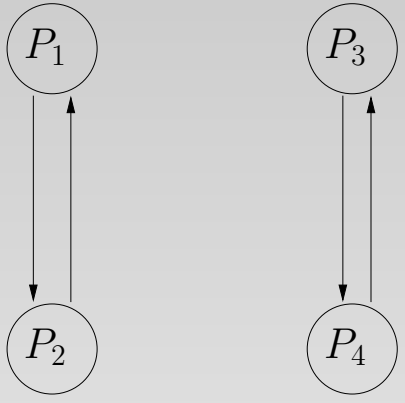
$$+\frac{1-\alpha}{n}\left(x_1+x_2+\cdots+x_n\right)$$

tilføjes højresiden.

- Samlet set giver det PageRank ligningerne for netværket.
- I googles PageRank er $\alpha \approx 0.85$

Google-ligningerne for eksemplet

For netværket



får vi

$$x_1 = \alpha x_2 + (1-\alpha)(x_1 + x_2 + x_3 + x_4)/4$$
$$x_2 = \alpha x_1 + (1-\alpha)(x_1 + x_2 + x_3 + x_4)/4$$
$$x_3 = \alpha x_4 + (1-\alpha)(x_1 + x_2 + x_3 + x_4)/4$$
$$x_4 = \alpha x_3 + (1-\alpha)(x_1 + x_2 + x_3 + x_4)/4$$

Ikke overraskende er den normaliserede løsning

$$x_1 = x_2 = x_3 = x_4 = \frac{1}{4}$$

Går det altid godt?

Hundred år gammel matematik giver et klart svar på spørgsmålet om eksistens og entydighed af en PageRank løsning:

Sætning: (Perron 1907):
Givet et ligningssystem opstået udfra et netværk som ovenfor beskrevet (med modifikationer til løsning af problem 1-2).
Da findes altid en entydig normaliseret og positiv løsning!

Hvordan beregnes løsningen?

- Kompleksiteten for løsning af ligningssystemer er $\approx n^2$ (n er antallet af ligninger). Dyrt når $n \approx 10^9$. (10^{18} operationer på en 3GHz maskine tager 4000 døgn!)
- Vi kan bruge en iterativ algoritme til at beregne løsningen: Givet $(x_1^{(0)}, x_2^{(0)}, x_2^{(0)}, x_3^{(0)})$ defineres rekursivt for $k \geq 1$

$$x_1^{(k)} = \alpha x_2^{(k-1)} + (1-\alpha)(x_1^{(k-1)} + x_2^{(k-1)} + x_3^{(k-1)} + x_4^{(k-1)})/4$$
$$x_2^{(k)} = \alpha x_1^{(k-1)} + (1-\alpha)(x_1^{(k-1)} + x_2^{(k-1)} + x_3^{(k-1)} + x_4^{(k-1)})/4$$
$$x_3^{(k)} = \alpha x_4^{(k-1)} + (1-\alpha)(x_1^{(k-1)} + x_2^{(k-1)} + x_3^{(k-1)} + x_4^{(k-1)})/4$$
$$x_4^{(k)} = \alpha x_3^{(k-1)} + (1-\alpha)(x_1^{(k-1)} + x_2^{(k-1)} + x_3^{(k-1)} + x_4^{(k-1)})/4.$$

  Da vil $(x_1^{(k)}, x_2^{(k)}, x_2^{(k)}, x_3^{(k)}) \to (x_1, x_2, x_2, x_3)$ når $k \to \infty$.
- I praksis er det tilstrækkeligt med $k = 50$.

Med mere...

Googles algoritme bygger på mange andre elementer end link-analyse, herunder

- Analyse af hjemmesidens titel, tekst (naturligt sprog)
- Site troværdighed (.edu og .gov står stærkere end .com)
- Alderen på hjemmesiden og sitet
- Omdømmet af de sider, der linker til hjemmesiden (mange link fra upålidelige kilder giver anledning til dårlig score)
- Blacklisting af kendte, utroværdige sider (link farms)
- Google sørger for at give et varieret svar på forespørgsler

Litteratur

- Kurt Bryan and Tanya Leise, *The $25.000.000.000 Eigenvector: The Linear Algebra behind Google*, SIAM Review **48** (3), 2006
- Amy N. Langville and Carl D. Meyer, *Google's PageRank and Beyond*, Princeton University Press, 2006
- D. Laksov, *Matematikk og Informasjonssøkning på nettet*, Normat **51** (3), 2003