Google Matrix & Problém relevance webových stránek

Martin Plešinger

Seminář KO-MIX, TU v Liberci

19. listopadu 2018

Širší kontext úlohy

V neustále rozrůstajícím se virtuálním prostoru internetu (resp. především webu) přirozená vzniká potřeba efektivně vyhledávat.

Na začátku 21. století přestávají na tuto práci stačit klasické rozcestníky (de-facto seznamy odkazů), je potřeba mít k dispozici efektivnější nástroj, který bude pro uživatele provádět navíc nějaký předvýběr na základě *relevance*.

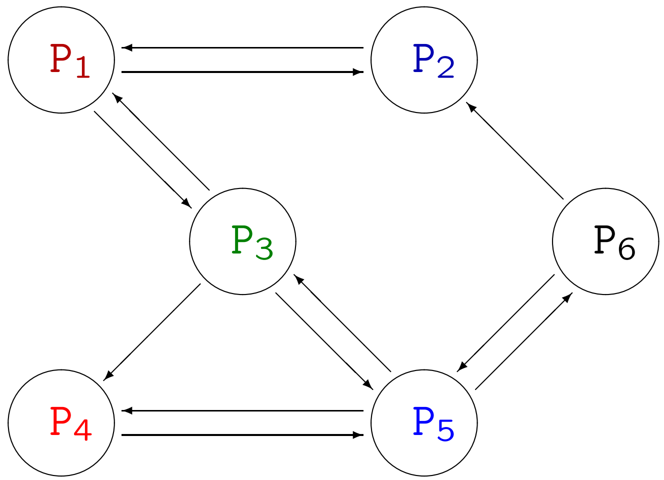
Základní pozorování, nástroje a předpoklady:

- Soubor dokumentů na webu budeme modelovat jako *graf*. Dokumenty na sebe odkazují jednosměrnými odkazy, *hyperlinky*, graf tedy bude *orientovaný*.
- O uživateli obecně nemáme žádnou a-priorní informaci, budeme tedy předpokládat, že se pohybuje po webu náhodně, tzv. *náhodný surfař*.
- Jeho pohyb po grafu popíšeme pomocí *pravděpodobností*. Tím z grafu vytvoříme tzv. *Markovův řetězec*.
- Klíčové budou vztahy mezi grafy, resp. markovovými řetězci a *maticemi sousednosti* resp. *přechodu*.

Viz např. [Langville, Meyer: 2006], [Berry, Browne: 2005].

1. Základní pohyb na webu

Uvažujme modelový web:



kde $\mathtt{P}_k$ je k-tý dokument na webu. Nechť dále
$\mathcal{B}_k$ je množina stránek odkazujících na $\mathtt{P}_k$ a
$\mathcal{F}_k$ je množina stránek, na které $\mathtt{P}_k$ odkazuje; např

$$\mathcal{B}_3 = \{\mathtt{P}_1, \mathtt{P}_5\}, \quad \mathcal{F}_5 = \{\mathtt{P}_3, \mathtt{P}_4, \mathtt{P}_6\}.$$

2. Co je to PageRank?

Larry Page* přichází s jednoduchou myšlenkou (algoritmem na) hodnocení stránek následně po něm pojmenovanou PageRank:

Relevanci stránky $\mathtt{P}_k$ definujme jako

$$r(\mathtt{P}_k) = \sum_{\mathtt{P}_j \in \mathcal{B}_k} \frac{r(\mathtt{P}_j)}{|\mathcal{F}_j|}, \qquad r(\mathtt{P}_k) > 0\,.$$

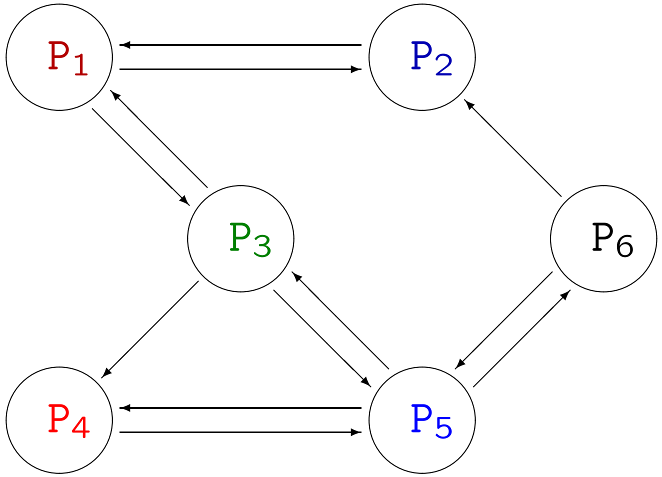
Interpretace:
(1.) Odkazuje-li na $\mathtt{P}_k$ důležitá stránka $\mathtt{P}_j$, zvyšuje tím důležitost $\mathtt{P}_k$.
(2.) Čím více stránek odkazuje na $\mathtt{P}_k$, tím je $\mathtt{P}_k$ důležitější.
(3.) Pokud nějaká stránka odkazuje na více stránek, její důležitost se mezi odkazované stránky rozdělí rovným dílem.

Jak si poradit se "zacyklenou" definicí (relevance je definována pomocí sebe sama)?

---

*Larry Page a Sergej Brin jsou zakladatelé společnosti Google.

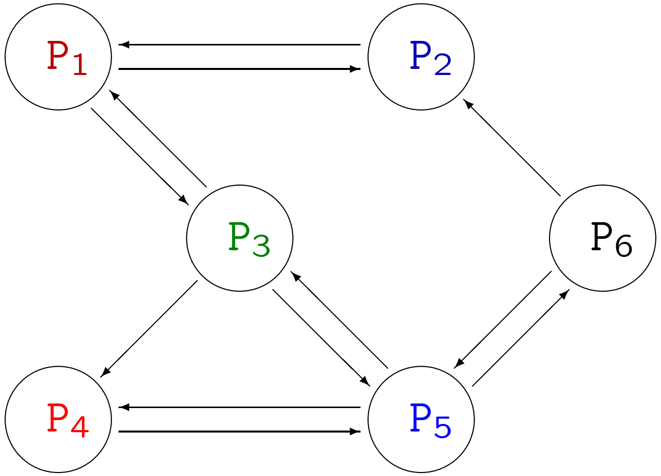
Pro náš modelový web



zřejmě platí:

$$r(\mathtt{P}_1) = r(\mathtt{P}_2) + \tfrac{1}{3} r(\mathtt{P}_3),$$

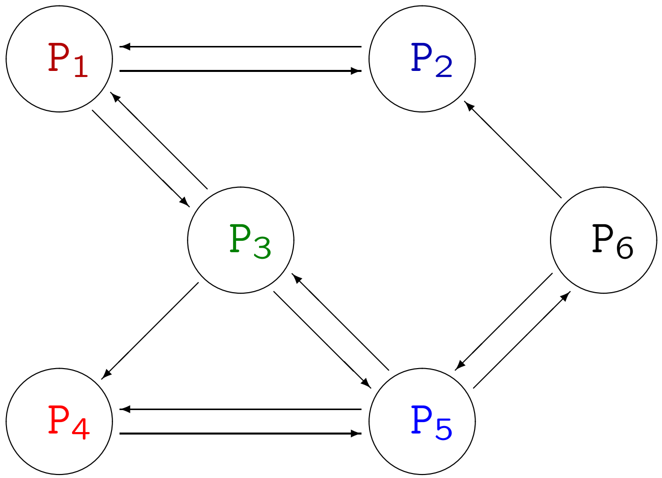
Pro náš modelový web



zřejmě platí:

$$
\begin{aligned}
r(\mathtt{P}_1) &= r(\mathtt{P}_2) + \tfrac{1}{3} r(\mathtt{P}_3), \\
r(\mathtt{P}_2) &= \tfrac{1}{2} r(\mathtt{P}_1) + \tfrac{1}{2} r(\mathtt{P}_6),
\end{aligned}
$$

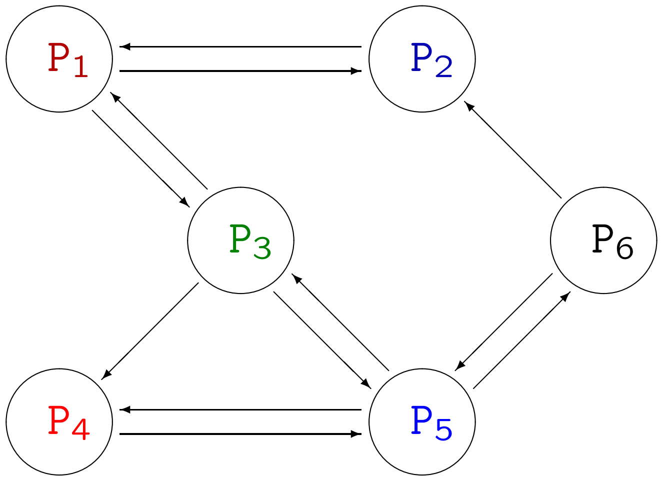
Pro náš modelový web



zřejmě platí:

$$
\begin{aligned}
r(\mathtt{P}_1) &= \qquad r(\mathtt{P}_2) + \tfrac{1}{3}r(\mathtt{P}_3),\\
r(\mathtt{P}_2) &= \tfrac{1}{2}r(\mathtt{P}_1) \qquad + \tfrac{1}{2}r(\mathtt{P}_6),\\
r(\mathtt{P}_3) &= \tfrac{1}{2}r(\mathtt{P}_1) \qquad + \tfrac{1}{3}r(\mathtt{P}_5),
\end{aligned}
$$

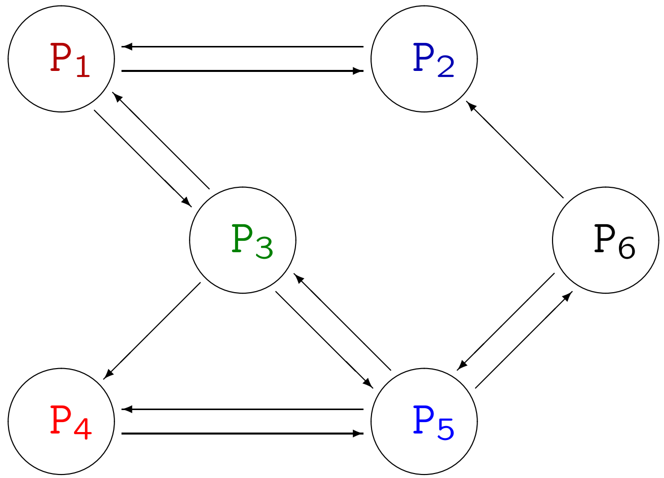
Pro náš modelový web



zřejmě platí:

$$
\begin{aligned}
r(P_1) &= r(P_2) + \tfrac{1}{3} r(P_3),\\
r(P_2) &= \tfrac{1}{2} r(P_1) + \tfrac{1}{2} r(P_6),\\
r(P_3) &= \tfrac{1}{2} r(P_1) + \tfrac{1}{3} r(P_5),\\
r(P_4) &= \tfrac{1}{3} r(P_3) + \tfrac{1}{3} r(P_5),
\end{aligned}
$$

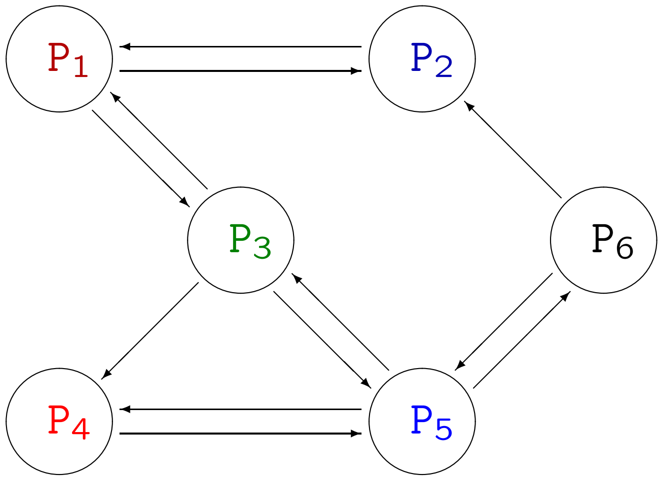
Pro náš modelový web



zřejmě platí:

$$
\begin{aligned}
r(P_1) &= && r(P_2) + \tfrac{1}{3} r(P_3), \\
r(P_2) &= \tfrac{1}{2} r(P_1) && + \tfrac{1}{2} r(P_6), \\
r(P_3) &= \tfrac{1}{2} r(P_1) && + \tfrac{1}{3} r(P_5), \\
r(P_4) &= && \tfrac{1}{3} r(P_3) + \tfrac{1}{3} r(P_5), \\
r(P_5) &= && \tfrac{1}{3} r(P_3) + r(P_4) + \tfrac{1}{2} r(P_6),
\end{aligned}
$$

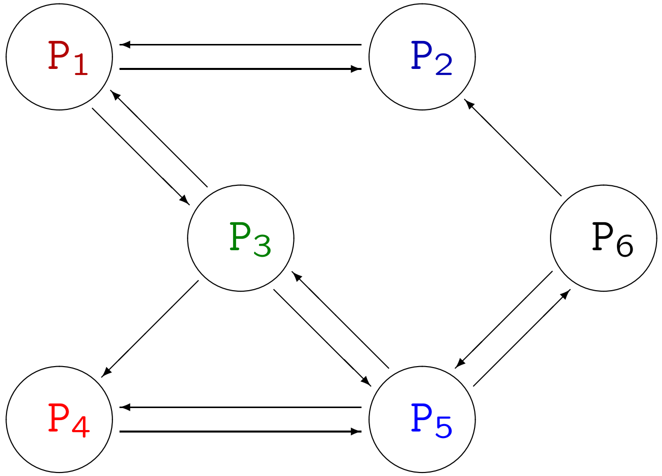
Pro náš modelový web



zřejmě platí:

$$
\begin{aligned}
r(\mathrm{P}_1) &= r(\mathrm{P}_2) + \tfrac{1}{3}r(\mathrm{P}_3),\\
r(\mathrm{P}_2) &= \tfrac{1}{2}r(\mathrm{P}_1) + \tfrac{1}{2}r(\mathrm{P}_6),\\
r(\mathrm{P}_3) &= \tfrac{1}{2}r(\mathrm{P}_1) + \tfrac{1}{3}r(\mathrm{P}_5),\\
r(\mathrm{P}_4) &= \tfrac{1}{3}r(\mathrm{P}_3) + \tfrac{1}{3}r(\mathrm{P}_5),\\
r(\mathrm{P}_5) &= \tfrac{1}{3}r(\mathrm{P}_3) + r(\mathrm{P}_4) + \tfrac{1}{2}r(\mathrm{P}_6),\\
r(\mathrm{P}_6) &= \tfrac{1}{3}r(\mathrm{P}_5).
\end{aligned}
$$

Pro náš modelový web



tedy dostáváme matici *hyperlinků* H:

$$\underbrace{\begin{bmatrix} r(P_1) \\ r(P_2) \\ r(P_3) \\ r(P_4) \\ r(P_5) \\ r(P_6) \end{bmatrix}}_{\pi} = \underbrace{\begin{bmatrix} 0 & 1 & 1/3 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 1/2 \\ 1/2 & 0 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 0 \\ 0 & 0 & 1/3 & 1 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/3 & 0 \end{bmatrix}}_{H^T} \underbrace{\begin{bmatrix} r(P_1) \\ r(P_2) \\ r(P_3) \\ r(P_4) \\ r(P_5) \\ r(P_6) \end{bmatrix}}_{\pi},$$

$$H^T\pi = \pi = \pi \cdot \lambda, \quad \lambda = 1,$$
$$\pi^T H = \pi^T.$$

3. Matice hyperlinků a její vlastnosti

Matice

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

• je po prvcích nezáporná, dokonce $0 \leq H \leq 1$, a součet prvků v řádku je roven jedné, tzv. stochastická matice.

• Vektor $e = [1, \ldots, 1]^T$ je zřejmě (pravým) vlastním vektorem, $He = e = e \cdot 1$ odpovídajícím vlastnímu číslu $\lambda = 1$.

• Hledaný vektor $\pi = [p(\mathtt{P}_1), \ldots p(\mathtt{P}_n)]^T$ je odp. levým vlastním vektorem, $\pi^T H = \pi^T$.

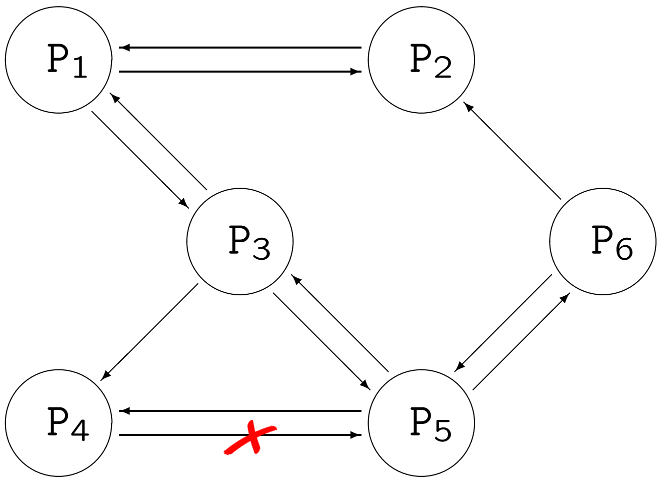
Prvky v j-tém řádku obsahují nuly a $\mathcal{F}_j$ čísel $1/|\mathcal{F}_j|$. Tedy $h_{j,k}$ lze interpretovat jako *pravděpodobnost přechodu* z $\mathtt{P}_j$ na $\mathtt{P}_k$ (náhodný surfař).

Orientovaný graf vybavený pravděpodobnostmi přechodu je *Markovovým řetězcem*.

Jsou tyto vlastnosti univerzální (platí to tak vždy)?

4. Teleportace

Pokud existuje stránka (dokument) $\mathtt{P}_j$, ze kterého nevede další odkaz, tj. $|\mathcal{F}_j| = 0$ (typicky obrázky `.jpg`, `.png`, ..., často i další dokumenty `.pdf`, atd.), angl. *dangling node*, je matice hyperlinků tzv. *substochastická*



$$H = \left[\begin{array}{cccccc} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1/3 & 1/3 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1/3 & 1/3 & 0 & 1/3 \\ 0 & 1/2 & 0 & 0 & 1/2 & 0 \end{array}\right]$$

Protože stochasticita je důležitá (nemáme vlastní pár $e = [1, \ldots, 1]^T$, $\lambda = 1$) a náhodný surfař by zůstal v takovém vrcholu grafu uvězněn, opravíme stochasticitu tím, že umožníme surfaři se *teleportovat* kamkoliv do grafu (v praxi uživatel zadá do adresního řádku jinou adresu).

Nechť $d = [d_1, \ldots, d_n]^T \in \{0,1\}^n$ (*dangling node vektor*) tak, že

$d_j = 1 \iff |\mathcal{F}_j| = 0 \iff \mathtt{P}_j$ je dangling node.

Pak *opravená matice hyperlinků* $H' = H + \frac{1}{n} d e^T$

$$= \left[\begin{array}{cccccc} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1/3 & 1/3 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1/3 & 1/3 & 0 & 1/3 \\ 0 & 1/2 & 0 & 0 & 1/2 & 0 \end{array}\right] + \frac{1}{6}\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$= \left[\begin{array}{cccccc} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1/3 & 1/3 & 0 \\ \hline 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ \hline 0 & 0 & 1/3 & 1/3 & 0 & 1/3 \\ 0 & 1/2 & 0 & 0 & 1/2 & 0 \end{array}\right]$$

je již vždy stochastická.

Vlastní vektor $\pi \in \mathbb{R}^n$ lze nalézt např. řešením singulární soustavy $(H' - I)\pi = 0$...

5. Spektrální poloměr (opravené) matice hyperlinků

Víme tedy, že $H'e = e$, tedy $\lambda = 1 \in \sigma(H')$. Je toto vlastní číslo něčím významné?

Nechť $v \in \mathbb{R}^n$, $\|v\|$ je libovolná vektorová norma, $A \in \mathbb{R}^{n \times n}$, pak

$$|||A||| = \max_{v,\ \|v\| = 1} \|Av\|$$

nazýváme operátorovou normou matice (indukovanou vektorovou normou $\|\cdot\|$).

Pro $Ax = x\lambda$, $\|x\| = 1$ zřejmě platí

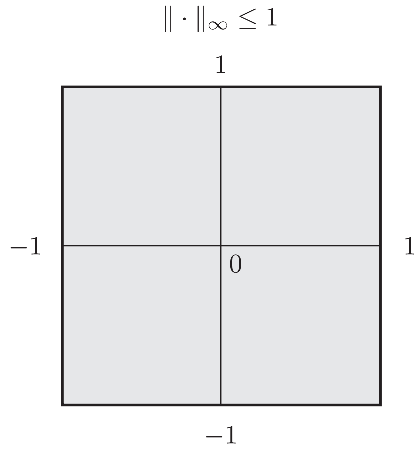
$$|||A||| = \max_{v,\ \|v\| = 1} \|Av\| \quad \geq \quad \|Ax\| = \|x\lambda\| = \|x\|\,|\lambda| = |\lambda|.$$

Označme spektrum a spektrální poloměr matice A

$$\sigma(A) = \{\lambda \in \mathbb{C} \,:\, \det(A - \lambda I) = 0\}, \qquad \varrho(A) = \max_{\lambda \in \sigma(A)} |\lambda|.$$

Tedy každá operátorová norma majorizuje spektrální poloměr matice

$$|||A||| \geq \varrho(A).$$



Označmě $v = [\nu_1, \ldots, \nu_n]^T$ a $A = (a_{j,k})$. Pak speciálně pro normu

$$\|v\|_\infty = \lim_{p\to\infty} \|v\|_p = \lim_{p\to\infty} \left(\sum_{j=1}^{n} |\nu_j|^p\right)^{\frac{1}{p}} = \max_j |\nu_j|$$

dostaneme

$$|||A|||_\infty = \max_j \sum_{k=1}^{n} |a_{j,k}| \qquad \text{(maximum ze součtů řádků } |A|\text{).}$$

Speciálně pro matici hyperlinků $H' = (h_{j,k})$, která je stochastická, dostaneme

$$|||H'|||_\infty = \max_j \sum_{k=1}^{n} |h_{j,k}| = 1.$$

Protože H' má vlastní číslo 1 a zároveň je $|||H'|||_\infty = 1 \geq \varrho(A) = \max_{\lambda\in\sigma(A)} |\lambda|$, je tedy $\lambda = 1$ *největší vlastní číslo* a zároveň *spektrální poloměr* matice H'.

Odpovídající vlastní vektor π umíme nalézt *mocninnou metodou*, pokud...

6. Mocninná metoda

Máme matici A a startovací vektor v_0, $\|v_0\| = 1$. Pro $\ell = 1, 2, 3, \ldots$ počítáme

$$p_\ell = A v_{\ell-1}, \quad v_\ell = \frac{p_\ell}{\|p_\ell\|} = \frac{1}{\|p_\ell\|} A\, v_{\ell-1} = \frac{1}{\varpi_\ell} A^\ell\, v_0,$$

kde $\varpi_\ell = \|p_\ell\| \cdot \cdots \cdot \|p_1\|$. Je-li A (pro jednoduchost) diagonalizovatelná, tj.

$$A = X \Lambda X^{-1}, \quad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n), \quad |\lambda_1| \geq \cdots \geq |\lambda_n|$$

pak pro $X = [x_1, \ldots, x_n]$ a $w_0 = X^{-1} v_0 = [\omega_1, \ldots, \omega_n]^T$

$$v_\ell = \frac{1}{\varpi_\ell} X \Lambda^\ell\, w_0 = \frac{1}{\varpi_\ell} \sum_{j=1}^{n} x_j \lambda_j^\ell\, \omega_j = \frac{\lambda_1^\ell}{\varpi_\ell} \left(x_1\, \omega_1 + \sum_{j=2}^{n} x_j \left(\frac{\lambda_j}{\lambda_1} \right)^\ell \omega_j \right).$$

Za jistých předpokladů ($\omega_1 \neq 0$, $|\lambda_1| > |\lambda_2|$ neboli $|\lambda_j / \lambda_1| < 1$) vidíme, že

$$v_\ell \sim x_1\, \varsigma_\ell, \quad \text{kde} \quad \varsigma_\ell = \frac{\lambda_1^\ell\, \omega_1}{\varpi_\ell}, \quad |\varsigma_\ell| \longrightarrow \frac{1}{\|x_1\|}, \quad \mathrm{span}(v_\ell) \longrightarrow \mathrm{span}(x_1).$$

7. Vlastní číslo $\lambda = 1$ může být násobné

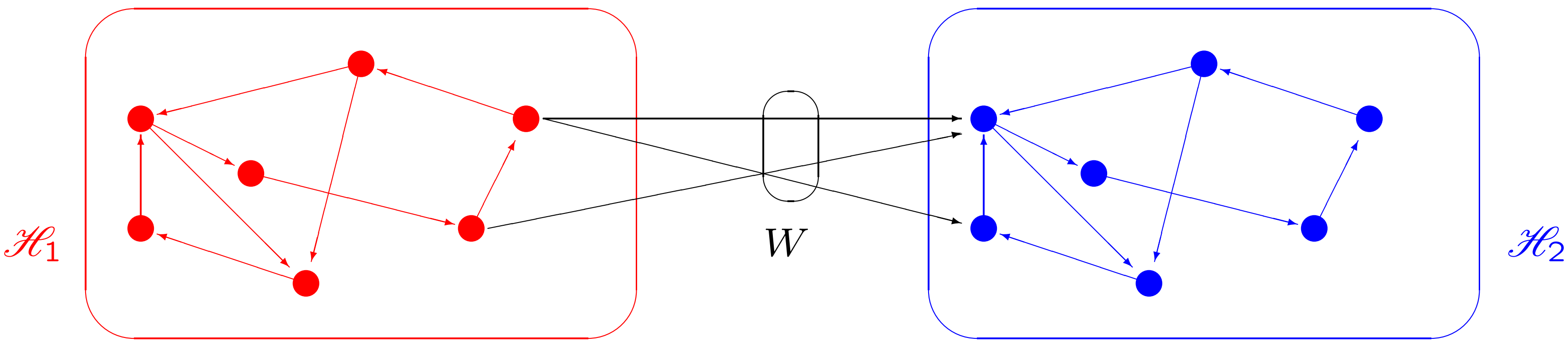
Např. když

$$H = \begin{bmatrix} H_1 & 0 \\ 0 & H_2 \end{bmatrix}, \quad \text{kde} \quad H_t \quad \text{jsou stochastické.}$$

Obecněji, když H je *rozložitelná*, tj. existuje permutační matice Π tak, že

$$H = \Pi \begin{bmatrix} H_1 & W \\ 0 & H_2 \end{bmatrix} \Pi^T, \quad \text{pak} \quad \sigma(H) = \sigma(H_1) \cup \sigma(H_2)$$

včetně násobností. Analogie dangling node, surfař zůstane uvězněn v H_2.



8. Aritmetika nezáporných a stochastických matic I.

Pro nezáporné matice zřejmě platí:

- Když $A \geq 0$, $B \geq 0$, pak $A + B \geq 0$.

Pro stochastické matice (čtvercové nezáporné s jednotkovými součty řádků) platí:

- Když $A \geq 0$ & $\sum_k a_{j,k} = 1$, $B \geq 0$ & $\sum_k b_{j,k} = 1$, pak

$$C = \alpha \cdot A + (1 - \alpha) \cdot B, \quad \sum_k c_{j,k} = \alpha \cdot \left(\sum_k a_{j,k} \right) + (1 - \alpha) \cdot \left(\sum_k b_{j,k} \right) = 1.$$

- Pokud navíc $0 \leq \alpha \leq 1$, pak $C \geq 0$.
- Pokud navíc $A > 0$ nebo $B > 0$ a $0 < \alpha < 1$, pak $C > 0$.

Tedy: *Konvexní kombinace stochastických matic je stochastická matice.*

9. Druhá teleportace

Náhodný surfař se může teleportovat kdykoliv, nejen v dangling nodes.

Zavedeme tzv. *Google Matrix*

$$G = \alpha \cdot \underbrace{\left(H + \frac{1}{n}\, de^T \right)}_{H'} + (1 - \alpha) \cdot \left(\frac{1}{n}\, ee^T \right), \qquad \text{kde} \qquad 0 < \alpha < 1$$

a čísla α a $(1 - \alpha)$ opět intepretujeme jako pravděpodobnosti.

G je stochastická, kladná a proto také nerozložitelná.

$\Rightarrow$ Náhodný surfař nyní nemůže uvíznout ve "slepé" ("visící") komponentě grafu.

$\Rightarrow$ Násobnost vlastního čísla $\lambda = 1$ je nyní "omezena" tím, že nemůže nastat situace ze slide 7. Stačí to ale?

10. Aritmetika nezáporných a stochastických matic II.

Pro nezáporné matice (a vektory) platí:

- [1] Když $A \geq 0$, $v \geq 0$, pak $Av \geq 0$.
- [2] Když $A \geq 0$, $B \geq 0$, pak $AB \geq 0$.
- [3] Když $A > 0$, $v \geq 0$ & $v \neq 0$, pak $Av > 0$.
- [4] Když $A > 0$, $B \geq 0$ & $B \neq 0$, pak $AB \geq 0$ & $AB \neq 0$.

10. Aritmetika nezáporných a stochastických matic II.

Pro nezáporné matice (a vektory) platí:

- $\bullet^1$ Když $A \geq 0$, $v \geq 0$, pak $Av \geq 0$.
- $\bullet^2$ Když $A \geq 0$, $B \geq 0$, pak $AB \geq 0$.
- $\bullet^3$ Když $A > 0$, $v \geq 0$ & $v \neq 0$, pak $Av > 0$.
- $\bullet^4$ Když $A > 0$, $B \geq 0$ & $B \neq 0$, pak $AB \geq 0$ & $AB \neq 0$.

V důsledku dále platí:

- $\bullet^5$ Když $A \geq 0$, $x \geq y$, pak $Ax \geq Ay$.

D: $x \geq y \;\Rightarrow\; (x - y) \geq 0 \;\Rightarrow^1\; A(x - y) \geq 0 \;\Rightarrow\; Ax - Ay \geq 0 \;\Rightarrow\; Ax \geq Ay$. $\square$

10. Aritmetika nezáporných a stochastických matic II.

Pro nezáporné matice (a vektory) platí:

- $\bullet^1$ Když $A \geq 0$, $v \geq 0$, pak $Av \geq 0$.
- $\bullet^2$ Když $A \geq 0$, $B \geq 0$, pak $AB \geq 0$.
- $\bullet^3$ Když $A > 0$, $v \geq 0$ & $v \neq 0$, pak $Av > 0$.
- $\bullet^4$ Když $A > 0$, $B \geq 0$ & $B \neq 0$, pak $AB \geq 0$ & $AB \neq 0$.

V důsledku dále platí:

- $\bullet^5$ Když $A \geq 0$, $x \geq y$, pak $Ax \geq Ay$.

D: $x \geq y \;\Rightarrow\; (x-y) \geq 0 \;\Rightarrow^1\; A(x-y) \geq 0 \;\Rightarrow\; Ax - Ay \geq 0 \;\Rightarrow\; Ax \geq Ay$. $\square$

- $\bullet^6$ Když $A > 0$, $x \geq y$ & $x \neq y$, pak $Ax > Ay$.

D: $x \geq y$ & $x \neq y \;\Rightarrow\; (x-y) \geq 0$ & $(x-y) \neq 0 \;\Rightarrow^3\; A(x-y) > 0 \;\Rightarrow\; Ax > Ay$. $\square$

•[7] Když $A \geq 0$, $v \geq 0$, $Av > v\,\varphi$, pak:

V první řadě $v \neq 0$. Pro $\varphi \geq 0$ navíc $w \equiv Av > 0$.

Označme $v = [\nu_1, \ldots, \nu_n]^T$, $w = [\omega_1, \ldots, \omega_n]^T$, pak $\omega_j > \nu_j\,\varphi$, $j = 1, \ldots, n$, a pro

$$\varepsilon = \min_s \frac{\omega_s}{\nu_s} - \varphi > 0, \quad \text{dostaneme} \quad Av \geq v\,(\varphi + \varepsilon).$$

Označme dále $B = \frac{1}{\varphi+\varepsilon} A$, platí tedy $B \geq 0$, $Bv \geq v$ a v důsledku[5] dostáváme

$$\cdots \geq B^{\ell+1}v \geq B^{\ell}v \geq B^{\ell-1}v \geq \cdots \geq B^2v \geq Bv \geq v.$$

Poznamenejme, že pro čtvercovou matici M platí

$$M^{\ell} \longrightarrow 0 \qquad \Leftrightarrow \qquad \varrho(M) < 1.$$

Zřejmě $B^{\ell} \not\longrightarrow 0$, tedy $\varrho(B) \geq 1$, tedy $\varrho(A) \geq \varphi + \varepsilon$, tedy

$$\varrho(A) > \varphi.$$

(Pro $\varphi < 0$ platí nerovnost triviálně.)

Nyní máme Google matrix $G > 0$ a PageRank vektor π tak, že $G^T\pi = \pi$, $\|\pi\|_\infty = 1$.

Pak $G^T|\pi| >^3 0$, přičemž $|\pi| \geq \pi$. Máme dvě možnosti:

1) $|\pi| = \pm\pi$ (všechny složky π mají stejná znaménka), pak

$$G^T|\pi| = \pm G^T\pi = |G^T\pi| = |\pi| \cdot 1.$$

2) $|\pi| \neq \pm\pi$ (vektor π má alespoň dvě nenulové složky různých znamének), pak

$$G^T|\pi| > |G^T\pi| = |\pi| \cdot 1, \qquad \text{a tedy} \qquad \varrho(G^T) = \varrho(G) >^7 1$$

což je spor s $\varrho(G) \leq \|G\|_\infty \max_{v,\ \|v\|_\infty = 1} \|Gv\|_\infty = 1$.

Důsledek: Vektor PageRanků $\pi = [r(\mathtt{P}_1), \ldots, r(\mathtt{P}_n)]^T$ lze volit nezáporný.

Protože $G > 0$, $\pi \geq 0$, $\pi \neq 0$ a $\pi \cdot 1 = G^T\pi$ dostáváme navíc

$$r(\mathtt{P}_j) = \sum_{k=1}^{n} g_{k,j}\, r(\mathtt{P}_k) > 0,$$

tedy vektor PageRanků lze vždy volit dokonce kladný $\pi > 0$.

11. Podruhé o násobnosti vlastního čísla $\lambda = 1$

Geometrická násobnost: Předpokládejme, že existují dva vlastní vektory

$$G^T v = v = [\nu_1, \dots, \nu_n]^T > 0, \quad G^T w = w = [\omega_1, \dots, \omega_n]^T > 0, \quad v \neq w.$$

Pak také $u = v\alpha + w\beta$ je vl. vektorem. Pro $\alpha = 1$ a $\beta = -\frac{\nu_j}{\omega_j}$ je j-tá složka u nulová, což je ve sporu s tím, že všechny jeho složky jsou stejného znaménka a nenulové.

- Geometrická násobnost vlastního čísla $\lambda = 1$ je tedy jedna.
- Matice G obsahuje pouze jeden Jordanův blok s tímto vlastním číslem.

11. Podruhé o násobnosti vlastního čísla $\lambda = 1$

Geometrická násobnost: Předpokládejme, že existují dva vlastní vektory

$$G^T v = v = [\nu_1, \ldots, \nu_n]^T > 0, \quad G^T w = w = [\omega_1, \ldots, \omega_n]^T > 0, \quad v \neq w.$$

Pak také $u = v\alpha + w\beta$ je vl. vektorem. Pro $\alpha = 1$ a $\beta = -\frac{\nu_j}{\omega_j}$ je j-tá složka u nulová, což je ve sporu s tím, že všechny jeho složky jsou stejného znaménka a nenulové.

- Geometrická násobnost vlastního čísla $\lambda = 1$ je tedy jedna.
- Matice G obsahuje pouze jeden Jordanův blok s tímto vlastním číslem.

Algebraická násobnost: Pravý a levý vlastní vektor netriviálního Jordanovu bloku

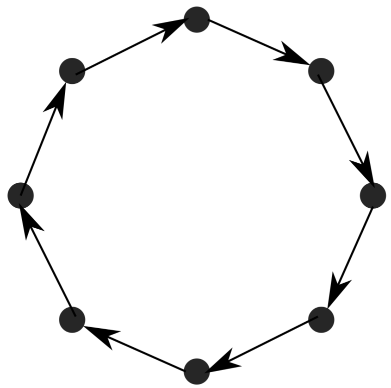
$$J = \begin{bmatrix} 1 & 1 & & & \\ & 1 & 1 & & \\ & & \ddots & \ddots & \\ & & & 1 & 1 \\ & & & & 1 \end{bmatrix}, \quad J v_{\mathfrak{R}} = v_{\mathfrak{R}} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \quad J^T v_{\mathfrak{L}} = v_{\mathfrak{L}} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

jsou navzájem vždy ortogonální $\langle v_{\mathfrak{R}}, v_{\mathfrak{L}} \rangle = 0$.

- Zřejmě $Ge = e = [1, \ldots, 1]^T$, $G^T \pi = \pi > 0$, $\langle e, \pi \rangle > 0$.
- Vlastní číslo $\lambda = 1$ matice G je jednoduché.

12. Poznámka o (im)primitivních maticích

Nemohou v matici existovat jiná vlastní čísla s absolutní hodnotou jedna?



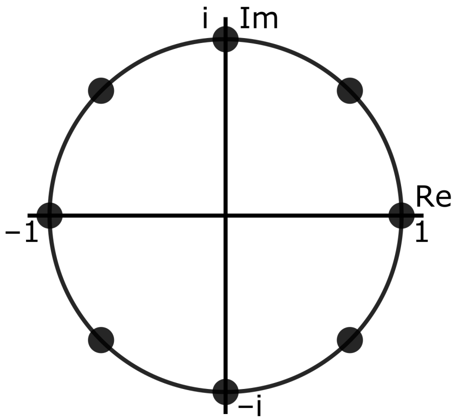
$$H = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\chi_H(\lambda) \\ = \det(\lambda I - H) \\ = \lambda^n - 1$$

Binomická rovnice $\lambda^n = 1$ však vede na $|\lambda_j| = 1$, $j = 1, \ldots, n$.

12. Poznámka o (im)primitivních maticích

Nemohou v matici existovat jiná vlastní čísla s absolutní hodnotou jedna?



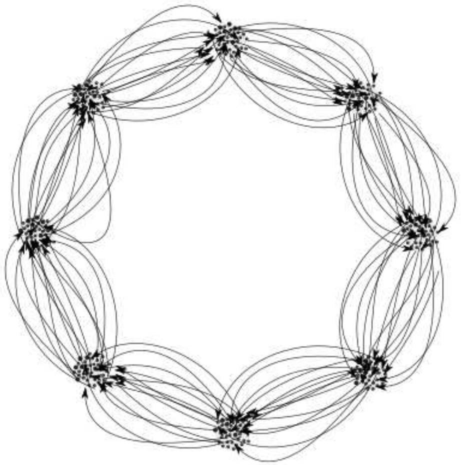
$$H = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\chi_H(\lambda) = \det(\lambda I - H) = \lambda^n - 1$$

Binomická rovnice $\lambda^n = 1$ však vede na $|\lambda_j| = 1$, $j = 1, \ldots, n$.

12. Poznámka o (im)primitivních maticích

Nemohou v matici existovat jiná vlastní čísla s absolutní hodnotou jedna?



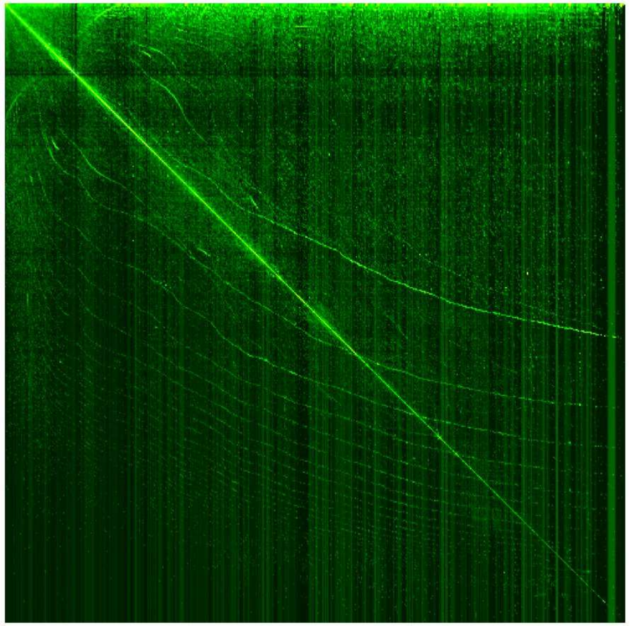
$$H = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & H_{1,n} \\ H_{2,1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & H_{3,2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & H_{4,3} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & H_{5,4} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & H_{6,5} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & H_{7,6} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & H_{8,7} & 0 \end{bmatrix}$$

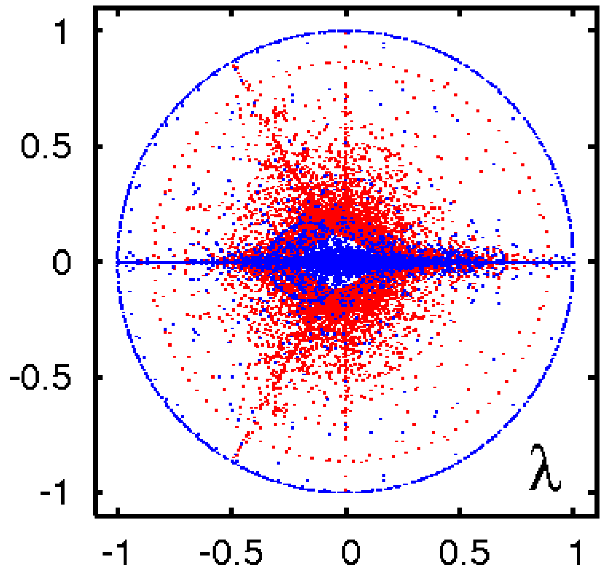
Délky všechny cyklů mají gcd $= 8 > 1$, tzv. *index imprimitivity.*

Nezáporná nerozložitelná matice s touto vlastností se nazývá *imprimitivní*.

Lze ukázat, že je-li matice kladná (tj. také Google Matrix G), pak už je vždy *primitivní*. Mocninná metoda tedy lze použít k nalezení PageRank vektoru π.

Několik faktů na závěr





Matice hyperlinků H stránek Cambridge University ($n = 212\,710$) a její spektrum.

- $n = 2.7 \times 10^9$ [Moler: 2002], 8.1×10^9 [LM: 2006], 11.5×10^9 [LSW: 2009].
- Matice je součtem řídké matice hyperlinků H, nnz$(H) \approx 10n$ [LM: 2006] a low-rank teleportační matice $\frac{1}{n}(d + e)e^T$; lze s ní manipulovat.
- Teleportace skrývá prostor pro personalizaci.

Reference

Amy N. Langville, Carl D. D. Meyer: *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, 2006

Michael W. Berry, Murray Browne: *Understanding Search Engines: Mathematical Modeling and Text Retrieval*, SE, SIAM, 2005

M. Fiedler: *Speciální matice a jejich použití v numerické matematice*, SNTL, 1981.

Cl. Moler: *The World's Largest Matrix Computation*, Matlab News & Notes, 2002.

Yiqin Lin, Xinghua Shi, Yimin Wei: *On computing PageRank via lumping the Google matrix*, JCAM 224(2) (2009), pp. 702–708.

Děkuji vám za pozornost