PageRank optimization applied to spam detection

Olivier Fercoq

The University of Edinburgh
Work completed while in INRIA Saclay and CMAP Ecole Polytechnique

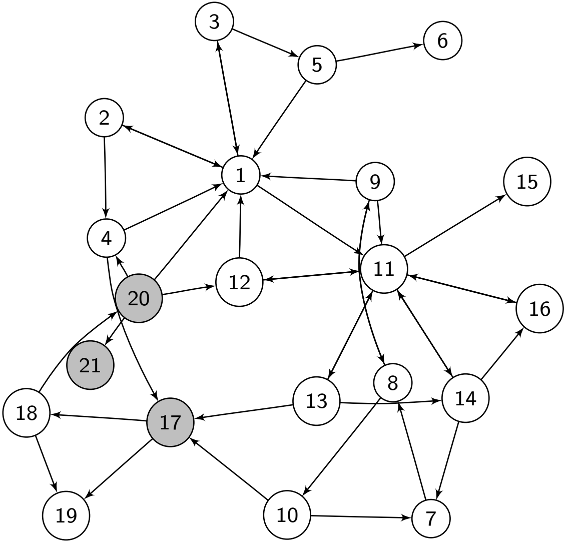
30th November 2012

Context

A webmaster controls a given number of pages:

- May add hyperlinks
- Must respect the content
  (the goal of a site is to provide information or service)
- Wishes to maximize:
  - Income (number of clicks on ads, number of sales)
  - Visibility (Sum of PageRank values of the site, PageRank of home page in Google)

Toy example with 21 pages



Nodes = web pages

Arcs = hyperlinks

21 : controlled page

1 : non controlled page

Definition of PageRank [Brin and Page, 1998]

- Random web surfer moves from page i to page j with probability $\frac{1}{D_i}$ ($D_i =$ degree of page i)
- $\pi =$ invariant measure of the Markov chain

$$\pi_i = \sum_{j:j\to i} \frac{\pi_j}{D_j}$$

- An important page is a page linked to by important pages
- Markov chain model may be reducible

Definition of PageRank [Brin and Page, 1998]

- Random web surfer moves from page i to page j with probability $\frac{1}{D_i}$ (D_i = degree of page i)
- π = invariant measure of the Markov chain

$$\pi_i = \alpha \sum_{j:j \to i} \frac{\pi_j}{D_j} + (1-\alpha) z_i$$

- An important page is a page linked to by important pages
- Markov chain model may be reducible
  $\to$ with probability $1-\alpha$, surfer gets bored and teleports: new research from page i with probability z_i
- Transition matrix: $P_{i,j} > 0, \forall i,j$ (usually $\alpha = 0.85$)
- PageRank is the unique invariant measure π of P

The PageRank optimization problem

- Well studied subject: Avratchenkov and Litvak, 2006
  Mathieu and Viennot 2006
  De Kerchove, Ninove and Van Dooren 2008
  Csáji, Jungers and Blondel 2010...
- Obligatory links $\mathcal{O}$, facultative links $\mathcal{F}$, prohibited links $\mathcal{I}$ (Strategy set proposed by Ishii and Tempo, 2010)
- Utility $\varphi(\pi, P) = \sum_i r_{i,j} \pi_i P_{i,j}$
- $r_{i,j}$ is viewed as reward by click on $i \to j$
- [Fercoq, Akian, Bouhtou, Gaubert, to appear in IEEE TAC]

Reduction to ergodic control

Proposition

$\mathcal{P}_i$ = set of admissible transition probabilities from Page i
The PageRank Optimization problem is equivalent
to the ergodic control problem with process X_t:

$$\max_{(\nu_t)_{t\geq 0}} \liminf_{T\to+\infty} \frac{1}{T}\mathbb{E}\Big(\sum_{t=0}^{T-1} r_{X_t,X_{t+1}}\Big)$$

$$\nu_t \in \mathcal{P}_{X_t}, \forall t \geq 0$$

$$\mathbb{P}(X_{t+1} = j | X_t = i, \nu_t = p) = p_j, \forall i,j \in [n], \forall p \in \mathcal{P}_i, \forall t \geq 0$$

where ν_t is a function of the history $(X_0, \nu_0, \ldots, X_{t-1}, \nu_{t-1}, X_t)$

Exponential size of the action sets

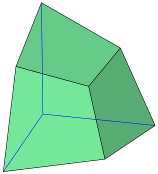
- At each page i, an action corresponds equivalently to
  - select $\nu \in \mathcal{P}_i$, a uniform measure on J
  - select $J \subseteq \mathcal{F}_i$
- 2^n hyperlink configurations by controlled page
- Classical Markov Decision Process techniques fail
- Csáji, Jungers and Blondel, 2010: graph rewriting to optimize the rank of a single page
- Our solution: action sets have a concise description

Admissible transition probabilities

Theorem

The convex hull of the set of admissible transition probabilities is either a simplex or a polyhedron defined by:

$$\begin{array}{ll} \forall j \in \mathcal{I}_i \ , & x_j = (1-\alpha) z_j \\ \forall j \in \mathcal{O}_i \setminus \{j_0\} \ , & x_j = x_{j_0} \\ \forall j \in \mathcal{F}_i \ , & (1-\alpha) z_j \leq x_j \leq x_{j_0} \\ \quad \textit{and} & \sum_{j \in [n]} x_j = 1 \end{array}$$



- Implicitly defined actions: vertices of the polytope
- Concise description ⇒ polynomial time separation oracle ⇒ well-described polyhedron [Groetschel, Lovász, Schrijver, 1988]

Well-described Markov Decision Processes

Define

A well-described MDP is a finite MDP where the action sets are defined *implicitly* as the vertices of well-described polyhedra (cf Groetschel, Lovász, Schrijver, 1988) and the transitions and rewards are linear

Theorem

The infinite horizon average cost problem on well-described MDP is solvable in polynomial time

Corollary

The PageRank optimization problem with local constraints is solvable in polynomial time

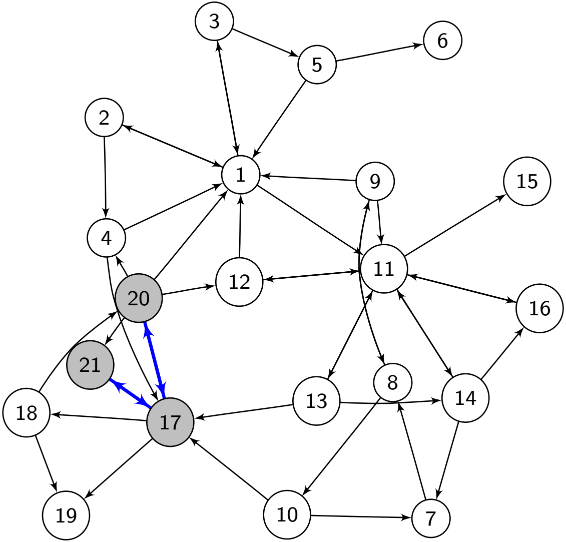
Resolution by Dynamic Programming

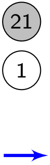
- The ergodic dynamic programming equation
$$w_i + \psi = \max_{\nu \in \mathcal{P}_i} \nu(r_{i,\cdot} + w), \quad \forall i \in [n] \tag{1}$$
has a solution $(w, \psi) \in \mathbb{R}^n \times \mathbb{R}$. The constant ψ is unique and is the value of the ergodic control problem
- To get an optimal strategy, select $\forall i$ a maximizing $\nu \in \mathcal{P}_i$

Resolution by Dynamic Programming

- The ergodic dynamic programming equation
$$w_i + \psi = \max_{\nu \in \mathcal{P}_i} \nu(r_{i,\cdot} + w), \quad \forall i \in [n] \tag{1}$$
has a solution $(w, \psi) \in \mathbb{R}^n \times \mathbb{R}$. The constant ψ is unique and is the value of the ergodic control problem
- To get an optimal strategy, select $\forall i$ a maximizing $\nu \in \mathcal{P}_i$
- The unique solution of the discounted equation
$$w_i = \max_{\nu \,:\, \alpha\nu + (1-\alpha)z \in \mathcal{P}_i} \alpha\nu(r_{i,\cdot} + w) + (1-\alpha)zr_{i,\cdot}, \forall i \in [n] \tag{2}$$
is solution of (1) with $\psi = (1-\alpha)zw$
- The fixed point scheme for (2) has contracting factor α independent of the dimension: complexity of optimization
$$\mathrm{O}\Big(\frac{\log(\epsilon)}{\log(\alpha)} \sum_{i \in [n]} |\mathcal{O}_i| + |\mathcal{F}_i| \log(|\mathcal{F}_i|)\Big)$$

Web graph optimized for PageRank

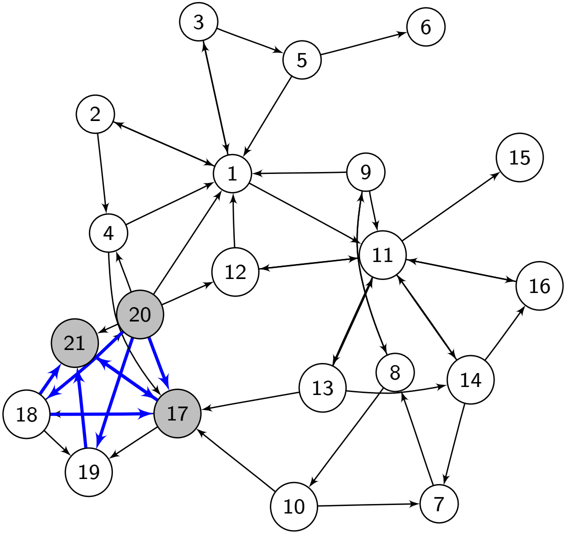




PageRank sum:
$0.10 \rightarrow 0.17$

The clique is not
an optimal startegy

Link spamming example



21 : spam web page

1 : honest page

18 : honeypot

added links

PageRank sum:

$0.10 \rightarrow 0.17 \rightarrow 0.31$

Search engine spamming

- Adding many unrelevant keywords
- Adding artificial pages that all point to a given page: Link farm [Gyöngyi and Garcia-Molina, 2005]
- Maximizing PageRank without design constraint [Baeza-Yates, Castillo and López, 2005]
- How to fight web spamming?

TrustRank and AntiTrustRank

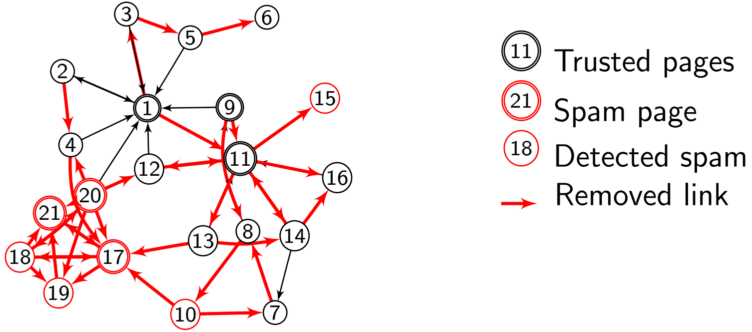
- Sets of hand-labelled trusted and spam pages
- Honest pages point to honest pages
- Spam pages are pointed to by spam pages
- TrustRank is a trust propagation algorithm: Compute PageRank with teleportation vector z such that $z_i > 0$ if and only if i is a trusted page. [Gyöngyi, Garcia-Molina, Pedersen, 2004]
- Distrust propagation with reversed hyperlinks: AntiTrustRank [Krishna and Raj, 2006]

Minimization of the PageRank of spam pages

- Trusted pages and known spam pages
- All the hyperlinks of the web are facultative
- Minimize the sum of PageRanks of spam pages

Minimization of the PageRank of spam pages

- Trusted pages and known spam pages
- All the hyperlinks of the web are facultative
- Minimize the sum of PageRanks of spam pages
- But no trust propagation



Penalty for hyperlink removals

- D_i hyperlinks in Page i in the original graph
- Selection of a set $J \in \mathcal{F}_i$ among the D_i hyperlinks
- A priori cost c'_i plus penalty for hyperlink removals ($\gamma > 0$)

$$c(i, J) = c'_i + \gamma \frac{D_i - |J|}{D_i}$$

- Additional control of teleportation vector:

$$z_j(I) = \begin{cases} 0 & \text{if } j \notin I \\ \frac{1}{N} & \text{if } j \in I \end{cases} \quad \text{for } I \subset [n], |I| = N < n$$

The MaxRank problem

Minimization of the PageRank of known spam pages with hyperlink removal penalty

$$\inf_{(I_t)_{t\geq 0},(J_t)_{t\geq 0}} \limsup_{T\to+\infty} \frac{1}{T}\mathbb{E}\Big(\sum_{t=0}^{T-1} c(X_t, J_t)\Big)$$

For all t, the currently visited page is X_t
The transitions are determined by:

$$I_t \subseteq [n], |I_t| = N \text{ and } J_t \subseteq \mathcal{F}_{X_t}$$

Well-described MDP formulation

$\mathcal{P}_i$ is the set of $(\sigma, \nu, w) \in \mathbb{R}^{D_i+1} \times \mathbb{R}^n$ such that

$$\begin{cases} \sum_{d=0}^{D_i} \sigma^d = 1 & \\ \sigma^d \geq 0 \ , & \forall d \in \{0, \dots, D_i\} \\ \nu_j = \sum_{d=0}^{D_i} w_j^d \ , & \forall j \in [n] \\ \sum_{j \in [n]} w_j^d = \sigma^d \ , & \forall d \in \{0, \dots, D_i\} \\ 0 \leq w_j^0 \leq \frac{\sigma^0}{N} \ , & \forall j \in [n] \\ w_j^d = 0 \ , & \forall j \notin \mathcal{F}_x, \forall d \in \{1, \dots, D_i\} \\ 0 \leq w_j^d \leq \frac{\sigma^d}{d} \ , & \forall j \in \mathcal{F}_x, \forall d \in \{1, \dots, D_i\} \end{cases}$$

$$\tilde{c}(i, \sigma, \nu, w) = c_i' + \gamma \frac{D_i - \sum_{d=0}^{D_i} d\sigma^d}{D_i},$$
$$\tilde{p}(y|i, \sigma, \nu, w) = \alpha \nu_y + (1 - \alpha) w_y^0$$

Fixed point operator

Proposition

Let T defined by

$$T_i(v) = \min_{(\sigma,\nu,w)\in\mathcal{P}_i} c_i' + \gamma \frac{D_i - \sum_{d=0}^{D_i} d\sigma^d}{D_i} + \alpha \sum_{j\in[n]} \nu_j v_j \ , \ \forall i \in [n]$$

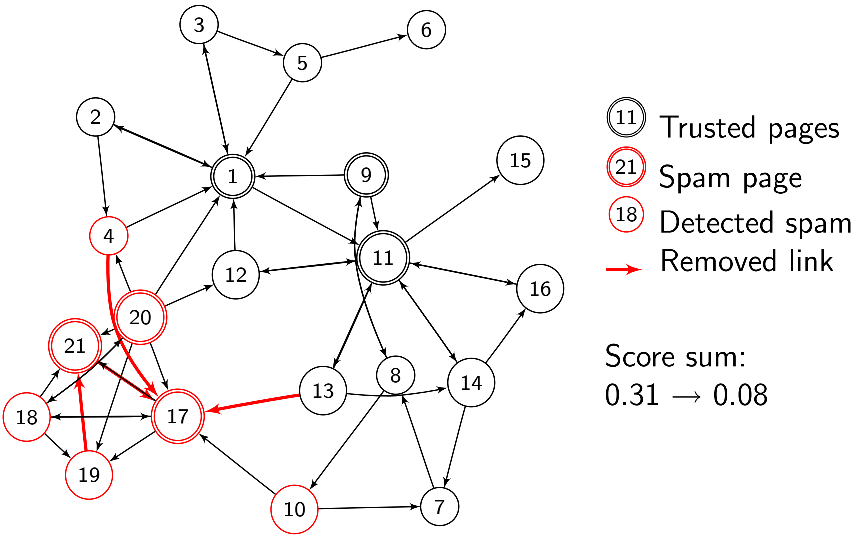
T is α-contracting with fixed point v

$(1-\alpha) \min_{w^0 \in Z} w^0 \cdot v$ is the value of the MaxRank problem

MaxRank bias

- The fixed point v is the bias of the ergodic control problem
- If $\gamma > \frac{2\alpha}{1-\alpha}\|c'\|_\infty$, then v_i is the expected mean number of spam pages visited before teleportation
  But no hyperlink is removed
- v_i gives a measure of the "spamicity" of Page i

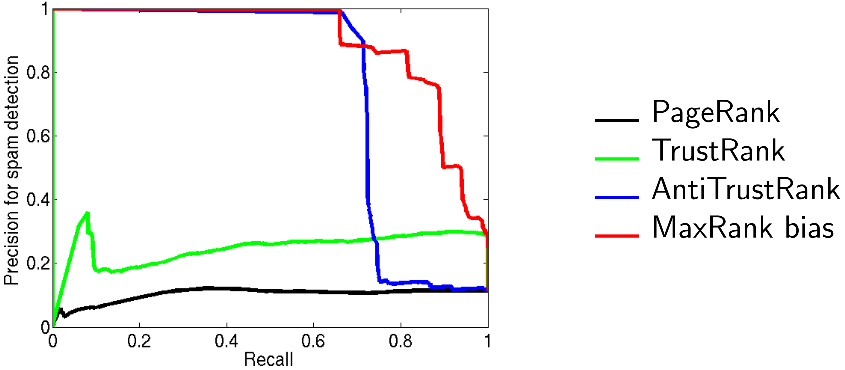
Toy example with $\gamma = 4$



Score sum:
$0.31 \rightarrow 0.08$

Spam detection by MaxRank bias

WEBSPAM-UK2007 dataset: 105,896,555 pages
Training set: 452,128 spam pages; 3,608,461 honest pages
Test set: 238,844 spam pages; 1,758,705 honest pages



Precision as a function of recall for spam detection

Conclusion

- Polynomial time solvability of the PageRank optimization problem
- Very fast optimization algorithm based on value iteration
- MaxRank: trust propagation algorithm based on PageRank optimization and well-described MDPs
- AUC $= 0.78$ within the range of WEBSPAM 2008 challengers $[0.73, 0.85]$