NCDawareRank

A Novel Ranking Method that Exploits the Decomposable Structure of the Web

Athanasios N. Nikolakopoulos   John D. Garofalakis

Computer Engineering and Informatics Department, University of Patras
Computer Technology Institute and Press "*Diophantus*"

Sixth ACM International Conference on Web Search and Data Mining
Rome 2013

Background

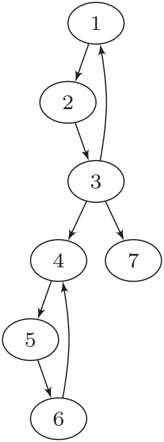
PageRank Model:

$$\mathbf{G} = \alpha \mathbf{H} + (1 - \alpha)\mathbf{E}$$

The **Damping Factor Issue**:

- Controls the fraction of importance, propagated through the links.
- The choice of α has received much attention
  - Picking very small $\alpha \Longrightarrow$ Uninformative Ranking Vector
  - Picking α close to $1 \Longrightarrow$ Computational Problems, Counterintuitive Ranking

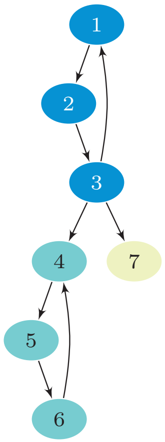
We focus on the **Teleportation model** itself!



Enriching the Teleportation Model

Web as a Nearly Completely Decomposable System:

- Nested Block Structure
  - Hierarchical Nature $\Longrightarrow$ **NCD Architecture**
- NCD has been exploited **Computationally**.
- We aim to exploit it **Qualitatively** in order to **Generalize the Teleportation Model**
  - Multiple Levels of Proximity between Nodes
  - **Core Idea**: Direct importance propagation to the NCD blocks that contain the outgoing links.



NCDawareRank Model I

$$\mathbf{P} = \eta\mathbf{H} + \mu\mathbf{M} + (1-\eta-\mu)\mathbf{E}$$

$$\mathbf{H} = [H_{uv}] \triangleq \frac{1}{d_u}, \quad \text{if } v \in \mathscr{G}_u$$

$$\mathbf{M} = [M_{uv}] \triangleq \frac{1}{N_u|\mathscr{A}_{(v)}|}, \quad \text{if } v \in \mathscr{X}_u$$

$$\text{where } \underbrace{\mathscr{X}_u \triangleq \bigcup_{w\in(u\cup\mathscr{G}_u)} \mathscr{A}_{(w)}}_{\text{Proximal Set of Pages}}$$

$$\mathbf{E} = \mathbf{e}\mathbf{v}^\intercal$$

- We partition the Web into **NCD blocks**, $\{\mathscr{A}_1, \mathscr{A}_2, \ldots, \mathscr{A}_N\}$,
- For every page u we define $\mathscr{X}_u$ to be its **proximal set** of pages, i.e the union of the NCD blocks that contain u and the pages it links to.
- We introduce an **Inter-Level Proximity Matrix M**, designed to propagate a fraction of importance to the proximal set of each page. Matrix **M** can be expressed as a product of 2 extremely sparse matrices, $\mathbf{R} \in \mathfrak{R}^{n\times N}$ and $\mathbf{A} \in \mathfrak{R}^{N\times n}$,
  - $\underbrace{n_z(\mathbf{R}) + n_z(\mathbf{A}) \ll n_z(\mathbf{H}) \ll n_z(\mathbf{M})}_{\textbf{efficient storage}}$
  - $\underbrace{\Omega_{\mathbf{R}\times\mathbf{A}} \ll \Omega_{\mathbf{H}} \ll \Omega_{\mathbf{M}}}_{\textbf{computability}}$

NCDawareRank Model II

Theorem (**Convergence Rate Bound:**)

The subdominant eigenvalue of matrix $\mathbf{P}$ *involved in the NCDawareRank, is upper bounded by* $\eta + \mu$.

Computational Experiments:

	PageRank	**NCDawareRank**						
	$\alpha = 0.85$	$\mu =$ 0.005	0.01	0.02	0.05	0.1	0.2	0.3
`cnr-2000`	48	47	45	43	41	**40**	40	41
`eu-2005`	42	42	41	40	39	**38**	40	41
`india-2004`	48	47	46	45	42	**42**	42	42
`indochina-2004`	47	46	45	44	42	**42**	42	42
`uk-2002`	46	45	44	43	42	**41**	41	41

Experimental Evaluation

Newly Added Pages Bias Problem:

- Methodology:
  - Extract the 90% of the incoming links of a set of randomly chosen pages.
  - Compare the orderings against those induced by the complete graph.

# New Pages	8000	10000	12000	15000	20000	30000
HyperRank	94.51±0.22	93.26±0.19	92.96±0.21	90.37±0.30	87.72±0.28	82.34±0.30
LinearRank	93.80±0.48	92.60±0.24	91.23±0.28	89.41±0.47	86.56±0.44	80.69±0.49
NCDawareRank	**96.81±1.06**	**96.48±1.10**	**96.64±0.42**	**95.44±1.39**	**94.77±0.72**	**91.49±1.42**
PageRank	93.68±0.59	92.46±0.30	91.04±0.37	89.19±0.55	86.33±0.53	80.26±0.57
RAPr	94.16±0.37	92.96±0.20	91.64±0.23	89.87±0.49	87.15±0.41	81.47±0.41
TotalRank	94.15±0.38	92.94±0.21	91.62±0.25	89.84±0.51	87.12±0.43	81.37±0.44

Sparsity:

- Methodology:
  - Randomly select to include 90% – 40% of the links on a new "sparsified" version of the graph
  - Compare the rankings of the algorithms against their corresponding original rankings.

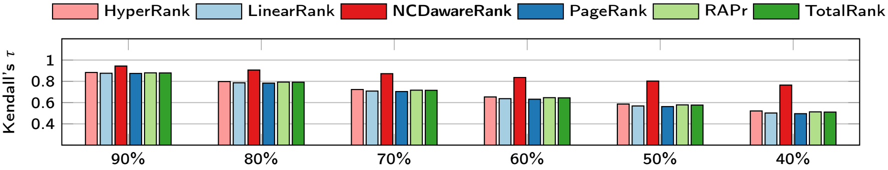
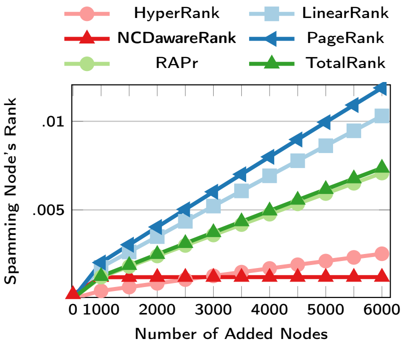


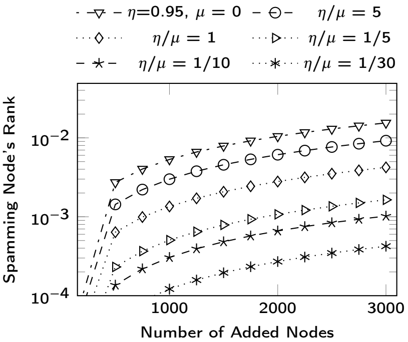
Fig 1. Ranking Stability under Sparseness.

Experimental Evaluation

Resistance to Direct Manipulation:

- Methodology:
  - Randomly pick a node with small initial ranking and we add a number of n nodes that funnel all their rank towards it.
  - We run all the algorithms for different values of n and we compare the spamming node's rank.





Conclussions and Future Research

We propose **NCDawareRank**:

- Generalizes PageRank by Enriching the Teleportation Model
- Produces More Stable Ranking Vectors
  - Sparseness Insensitivity
  - Resistance to Manipulation
- Opens new interesting research directions

Thanks!

Q&A