Ranking nodes in growing networks: When PageRank fails

Pietro De Nicolao
pietro.denicolao@mail.polimi.it

Politecnico di Milano

April 14, 2016

Outline

Outline

Paper

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Ranking nodes in growing networks: When PageRank fails

Manuel Sebastian Mariani[1], Matúš Medo[1] & Yi-Cheng Zhang[1,2]

Published on *Nature Scientific Reports* on 10 November 2015.

PageRank: recap

- **Most popular ranking algorithm for unipartite directed networks.**
- Invented for Google's search algorithm
- Also used for the ranking of:
  - scholarly papers
  - images in search
  - urban roads according to traffic flow
  - proteins in their interaction network
  - etc.
- **A node is important if it is pointed by other important nodes.**

$$p_{ij} = (1-\gamma)\frac{w_{ij}}{s_i^{out}} + \gamma\frac{1}{N}$$

One PageRank fits all?

What is the relation between PageRank's efficacy and the properties of the network?

- PageRank: static approach
  - PageRank discards temporal information
  - works as if nodes appear all at the same time
  - well-known bias towards old nodes
- Theoretical models and real networks can exhibit strong temporal patterns.

Are there circumstances under which the algorithm is doomed to fail?

Outline

The Relevance Model (RM)

- What is the Relevance Model?
  - growing directed network model
    with preferential attachment and relevance
  - generalizes the classical **Barabási-Albert** model
  - introduced by [Medo, 2011]
- Why using RM?
  - model that best explains the linking patterns in real networks
  - used to model WWW, citation and technological networks

Relevance Model features

1. Preferential attachment
   - similar to the Barabási-Albert model
   - Matthew effect: the rich get richer
   - significant difference: *existing nodes also create new links*
2. Fitness
   - quality parameter assigned to each node
   - *node's inherent competence in attracting new incoming links*
   - concept formerly explored in [Bianconi-Barabási, 2001]
3. Relevance and activity
   - **Relevance**: capacity of attracting new links over time
   - **Activity**: rate at which the node generates new outgoing links
4. Temporal decay
   - Relevance and activity both decay with time
   - Monotonous function of choice (exponential, power law)
   - Real-world phenomenon: nodes lose relevance over time

Relevance decay: a real-world example

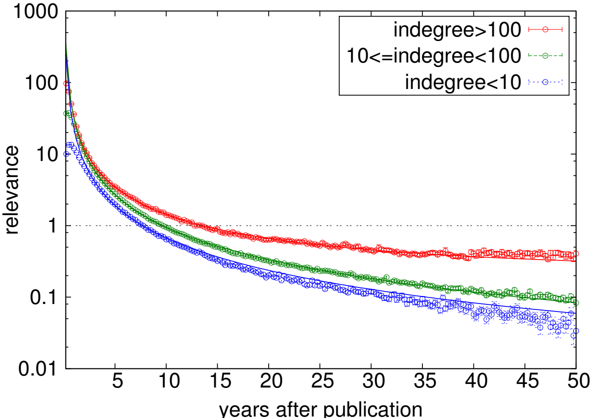


Figure: Temporal decay of the average (empirical) relevance $r(t)$ of papers in the American Physical Society citation network (1893-2009). This behaviour has been formerly highlighted in [Medo, 2011].

How to build a network with the Relevance Model

At each discrete time interval t, the **generation algorithm** proceeds as follows:

1. a new node is created and connected to an existing node i, chosen with probability $\Pi_i^{in}(t)$.
2. If $t > 10$, then $m = 10$ existing nodes are sequentially chosen with probability $\Pi_i^{out}(t)$ and become active:
    - each selected node creates one outgoing link
    - it selects a node j as a target with probability $\Pi_j^{in}(t)$

- No multiple links.
- No self loops.

Relevance Model: Link generation mechanism

The probability for the node i to be the target of a new link created at time t is:

$$\Pi_i^{in}(t) \sim (k_i^{in}(t) + 1)\, \eta_i \, f_R(t - \tau_i)$$

- $k_i^{in}(t)$: current indegree of node i
- η_i: fitness of i
- τ_i: time at which i enters the network
- f_R: monotonously decaying function of time
- $R_i(t) := \eta_i f_R(t - \tau_i)$: relevance of node i at time t.

Relevance Model: Active nodes selection

In the RM, nodes continue being active and generate outgoing links continually.
Probability for node i to be chosen as an **active node** at time t:

$$\Pi_i^{out}(t) \sim A_i \, f_A(t - \tau_i)$$

- A_i: activity parameter
- τ_i: time at which i enters the network
- f_A: monotonously decaying function of time

Effects of relevance decay in the RM

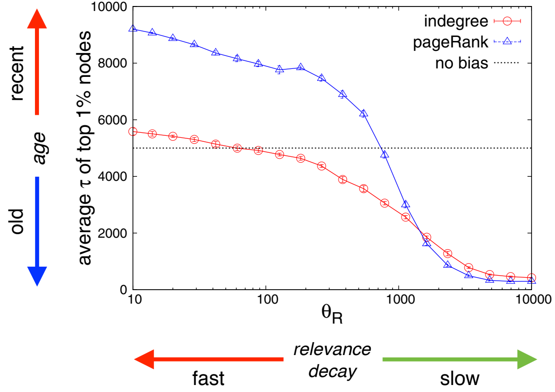
- Slow or absent relevance decay
  - recent nodes receive few links because of preferential attachment
  - PageRank's bias towards old nodes in scale-free networks
- Fast relevance decay
  - preferential attachment compensated by decay of relevance of old nodes
  - recent nodes can reach high indegree
  - recent nodes mostly point to other recent nodes, because of relevance decay of older nodes
  - old nodes point to nodes of every age because of activity

What makes a ranking algorithm "good"?

A good ranking algorithm is expected to produce an unbiased ranking where both recent and old nodes have the same chance to appear at the top.

- In growing networks with temporal effects, PageRank can fail to achieve this.
- Let's compare PageRank with the elementary indegree ranking.

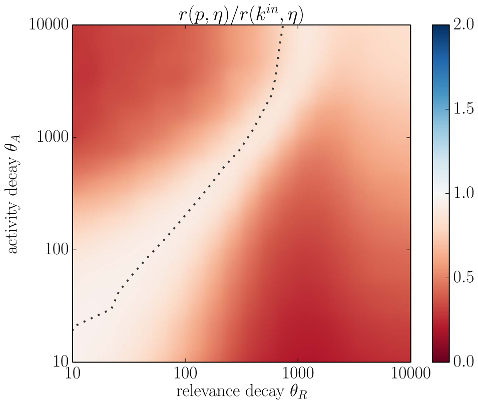
PageRank time bias: numerical simulation of RM



- Relevance decays as $f_R(t) = \exp(-\frac{t}{\theta_R})$.
- Activity decays exponentially, but very slowly ($\theta_A = N$).
- $N = 10000$

Figure: Average time of entrance of 1% of nodes of PageRank and indegree rankings, in the RM model.

PageRank vs. indegree: correlation with fitness in RM



- $r(p, \eta)$: correlation PageRank-fitness
- $r(k^{in}, \eta)$: correlation indegree-fitness
- $\rho(\eta) = \exp(-\eta)$
- $\rho(A) = 2A^{-3},\ A \in [1, \infty]$

Figure: Comparison of performance of PageRank and indegree (RM data). PageRank yields no improvement with respect to indegree. Diagonal: no temporal bias towards recent or old nodes.

Outline

Real networks studied in the paper

Real networks studied (directed, unweighted):

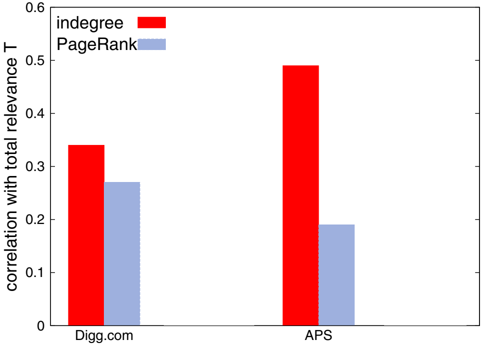
- Digg.com: social bookmarking site
  - Nodes: Digg users
  - Edges: $a_{ij} = 1 \Leftrightarrow$ "i is a follower of j".
  - $N = 190\,553$; $L = 1\,552\,905$
- American Physical Society (APS) articles and citation network
  - Nodes: papers (from 1893 to 2009)
  - Edges: $a_{ij} = 1 \Leftrightarrow$ "i cites j"
  - $N = 450\,056$; $L = 4\,690\,967$

Estimator of node fitness: total relevance of node i

$$T_i = \sum_t r_i(t)$$

To validate hypothesis of relevance and activity decay: measurement of empirical relevance (see appendix).

PageRank performance on real data



- Digg.com: activity and relevance decay s.t. PageRank is maximally correlated with indegree in RM simulations with power-law decay.
- APS: activity decays immediately, relevance decays progressively.

Figure: Comparison of PageRank and indegree correlation with total relevance T_i in real data. APS: PageRank strongly biased towards old nodes, because papers can only be cited by more recent papers.

Outline

Important findings

- PageRank can underperform w.r.t. indegree ranking
- Mismatch between relevance and activity decay timescales leads to time bias in PageRank:
  - towards recent nodes if decay of relevance is faster
  - towards old nodes if decay of activity is faster
- Findings are robust with respect to:
  - form of decay function
  - distribution of fitness among the nodes
  - metric used to evaluate the algorithm
- Link timestamps are crucial for this analysis
- Method can not be applied to undirected networks

In conclusion...

PageRank, despite its popularity and robustness, can fail and thus it should not be used without carefully considering the temporal properties of the system to which it is to be applied.

Bibliography

Mariani M. S., Medo M., Zhang Y.
Ranking nodes in growing networks: When PageRank fails.
Scientific Reports 5, 16181;
doi: 10.1038/srep16181 (2015).

G. Bianconi, A. L. Barabási
Competition and Multiscaling in evolving networks
Europhysics Letters, Vol. 54 (2001), pp. 436-442
doi:10.1209/epl/i2001-00260-6

Medo M., Cimini G., Gualdi S.
Temporal Effects in the Growth of Networks
Phys. Rev. Lett. 107, 238701 (2011-12-01)

Outline

Empirical relevance

The Extended Fitness Model

Empirical relevance: definition

The empirical relevance $r_i(t)$ of node i at time t is defined as:

$$r_i(t) = \frac{n_i(t)}{n_i^{PA}(t)}$$

- $n_i(t) = \frac{\Delta k_i^{in}(t,\Delta t)}{L(t,\Delta t)}$: ratio between:
  - $\Delta k_i^{in}(t, \Delta t)$: # of incoming links received by node i in the time window $[t, t + \Delta t]$
  - $L(t, \Delta t)$: total # of links created within the same time window
- $n_i^{PA}(t) = \frac{k_i^{in}(t)}{\sum_j k_j^{in}(t)}$: expected value of $n_i(t)$ according to preferential attachment alone

$r_i(t) > 1$ (< 1): node i at time t outperforms (underperforms) in the competition for incoming links with respect to its preferencial attachment weight.

Outline

Empirical relevance

The Extended Fitness Model

The Extended Fitness Model

- **PageRank's under-performance in time-dependent networks is a general feature.**
- We can validate this using a model more compatible with the idea that *a node is important if it's pointed by other important nodes.*

Extended Fitness Model (EFM)

- **High-fitness nodes are more sensitive to fitness** than low-fitness nodes, when choosing their outgoing links.
- High-fitness nodes are then more likely to be pointed by other high-fitness nodes than low-fitness nodes.
- EFM is more favorable to PageRank than RM.

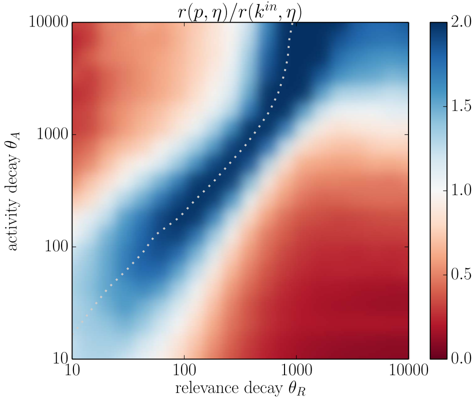
EFM: sensitivity to fitness

Probability $\Pi^{in}_{i;j}(t)$ that a link created by node j at time t ends in node i:

$$\Pi^{in}_{i;j}(t) \sim (k^{in}_i(t)+1)^{1-\eta_j}\,\eta_i^{\eta_j}\,f_R(t-\tau_i)$$

- Fitness $\eta \in [0,1]$ to prevent negative exponents
- Π^{in} depends on the fitness of the target *and of the source* nodes (difference with RM).
- $k^{in}_i(t)$: indegree of node i at time t.

PageRank vs. indegree: correlation with fitness in EFM



- $\rho(A) = 2A^{-3},\ A \in [1, \infty]$
- H nodes: high fitness; $\eta \in [10^{-5}, 1]$
- $(N - H)$ nodes: low fitness; $\eta \in [0, 10^{-5}]$
- $N = 10000,\ H = 250$

Figure: Comparison of performance of PageRank and indegree (EFM data).