Network metrics for reputation and quality in scholarly data

Matúš Medo

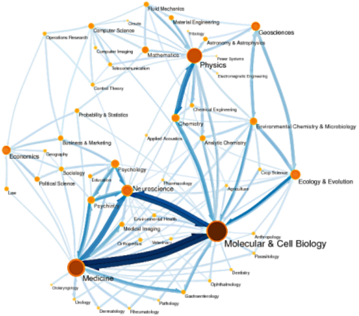
University of Fribourg, Switzerland

Symposium "Scientometric Maps and Dynamic Models of Science and Scientific Collaboration Networks"

10 March 2016, Regensburg

Part 1

Network-driven reputation
in online scientific communities

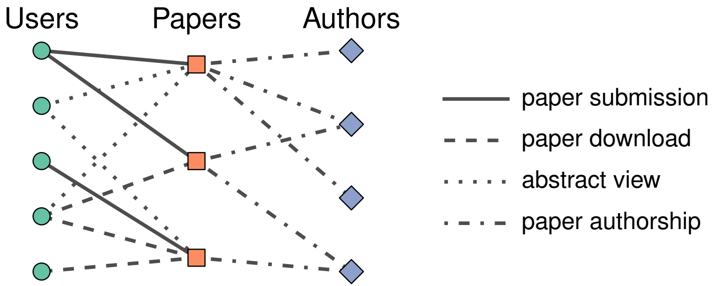


The setting

- Econophysics Forum: a web site for researchers and practitioners in econophysics and finance (`www.unifr.ch/econophysics`)
- Weblog data collected from 6th July 2010 to 31st March 2013 (1000 days in total)
- After data cleaning: 5071 users, 844 papers, 29 748 links

The setting

- Econophysics Forum: a web site for researchers and practitioners in econophysics and finance (`www.unifr.ch/econophysics`)
- Weblog data collected from 6th July 2010 to 31st March 2013 (1000 days in total)
- After data cleaning: 5071 users, 844 papers, 29 748 links

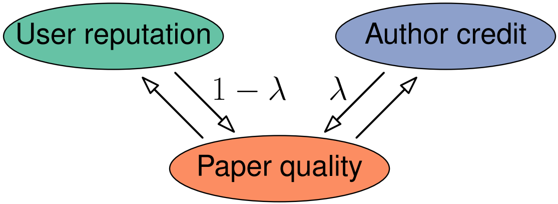


The basic idea

- Goal: to estimate paper quality from the feedback the paper has among the users
- But: papers also have authors—take their credit into account too
- In summary: user reputation R, paper quality/fitness F, and author credit A mutually depend on each other
  - Similar approach: PageRank, HITS, bipartite HITS, ...

The basic idea

- Goal: to estimate paper quality from the feedback the paper has among the users
- But: papers also have authors—take their credit into account too
- In summary: user reputation R, paper quality/fitness F, and author credit A mutually depend on each other
  - Similar approach: PageRank, HITS, bipartite HITS, ...



The QRC algorithm (Liao et al, 2014)

Users: $$R_i = \frac{1}{k_i^{\theta_R}} \sum_{\alpha=1}^{M} w_{i\alpha}(F_\alpha - \rho_F \bar{F}) \tag{1}$$

Authors: $$A_m = \frac{1}{d_m^{\phi_A}} \sum_{\alpha=1}^{M} P_{m\alpha}(F_\alpha - \rho_A \bar{F}) \tag{2}$$

Papers: $$F_\alpha = \frac{1-\lambda}{k_\alpha^{\theta_F}} \sum_{i=1}^{N} w_{i\alpha}(R_i - \rho_R \bar{R}) + \frac{\lambda}{d_\alpha^{\phi_P}} \sum_{m=1}^{O} P_{m\alpha} A_m \tag{3}$$

- Here:
  - $w_{i\alpha}$: weights of user-paper connections
  - $P_{m\alpha}$: paper authorship

The QRC algorithm (Liao et al, 2014)

Users: $$R_i = \frac{1}{k_i^{\theta_R}} \sum_{\alpha=1}^{M} w_{i\alpha}(F_\alpha - \rho_F \bar{F}) \tag{1}$$

Authors: $$A_m = \frac{1}{d_m^{\phi_A}} \sum_{\alpha=1}^{M} P_{m\alpha}(F_\alpha - \rho_A \bar{F}) \tag{2}$$

Papers: $$F_\alpha = \frac{1-\lambda}{k_\alpha^{\theta_F}} \sum_{i=1}^{N} w_{i\alpha}(R_i - \rho_R \bar{R}) + \frac{\lambda}{d_\alpha^{\phi_P}} \sum_{m=1}^{O} P_{m\alpha} A_m \tag{3}$$

- Here:
  - $w_{i\alpha}$: weights of user-paper connections
  - $P_{m\alpha}$: paper authorship
- $\rho_F, \rho_A, \rho_R > 0$: punishment for connections with low-rated nodes
- $\theta_R, \theta_F, \phi_A, \phi_P$: they decide whether we cumulate or average

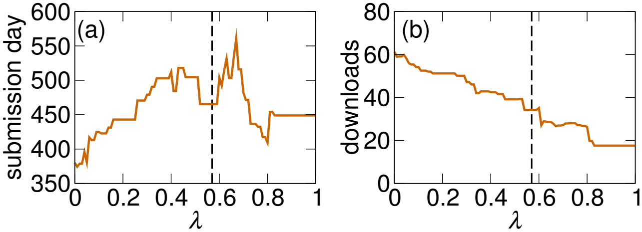
Context & the parametrization

- This is similar to Kleinberg's famous HITS, only with three layers
- Even more similar: Eigenrumor (Fujimura and Tanimoto, 2005) which has three layers but only two scores and different normalization

Context & the parametrization

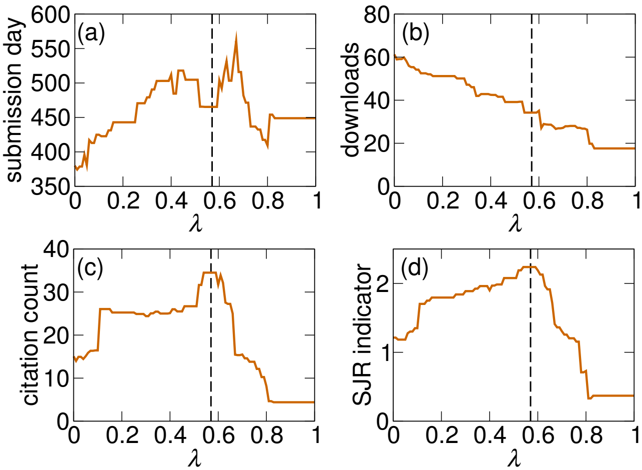
- This is similar to Kleinberg's famous HITS, only with three layers
- Even more similar: Eigenrumor (Fujimura and Tanimoto, 2005) which has three layers but only two scores and different normalization
- Our choice of parameters (motivated by artificial simulations and common sense)
  - $\theta_Q = 0$ (paper quality is a sum over all users who collect it)
  - $\theta_R = 1$ (user reputation is an average over all collected papers)
  - $\rho_F = \rho_R = \rho_A = 0$ (no penalty for connections with bad nodes)
  - $\phi_A = 0$ (author credit is a sum over all authored papers)
  - $\phi_P = 1$ (the average author credit contributes to paper quality)

Analysis of the top 20 papers



$\lambda = 0$: author credit has no impact on paper quality

Analysis of the top 20 papers



$\lambda = 0$: author credit has no impact on paper quality

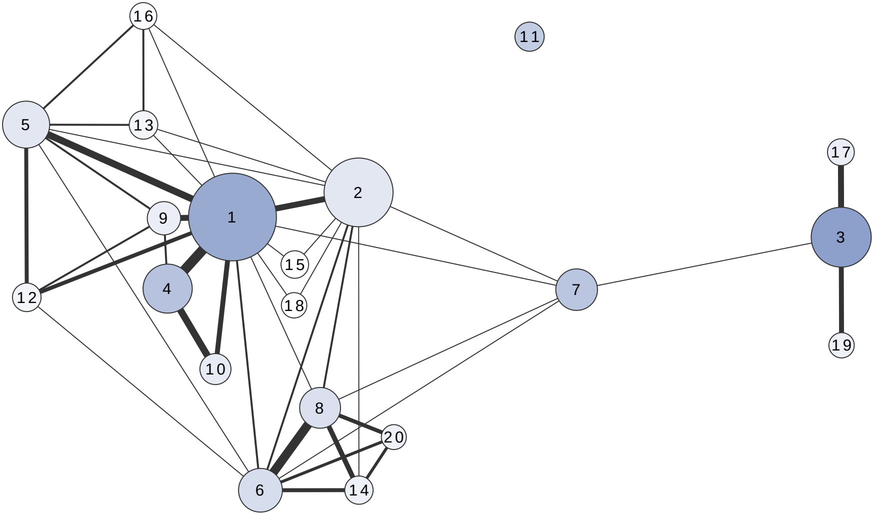
Analysis of the top 20 papers

Method	Day	Down	Cit	SJR
random	548 ± 41	11 ± 1	5 ± 1	0.5 ± 0.1
POP	299 ± 37	69 ± 7	15 ± 4	0.9 ± 0.4
biHITS	264 ± 34	56 ± 7	10 ± 3	0.7 ± 0.2
Eigenrumor	444 ± 49	30 ± 10	18 ± 4	0.9 ± 0.1
QR1	375 ± 49	59 ± 9	15 ± 4	1.2 ± 0.5
QR2	445 ± 47	54 ± 9	14 ± 3	1.2 ± 0.4
QRC	465 ± 60	34 ± 8	34 ± 10	2.2 ± 0.5

Analysis of the top 20 papers

Rank	Name	Credit	Papers	Down
1	H. E. Stanley	0.65	26	22
2	T. Preis	0.39	8	38
3	D. Sornette	0.35	29	17
4	S. Havlin	0.22	19	11
5	B. Podobnik	0.19	8	21
6	D. Y. Kenett	0.16	11	14
7	D. Helbing	0.16	18	20
8	E. Ben-Jacob	0.14	10	12
9	A. M. Petersen	0.10	6	13
10	S. V. Buldyrev	0.09	7	13
11	J.-P. Bouchaud	0.08	16	19
⋮	⋮	⋮	⋮	⋮
15	J. J. Schneider	0.07	1	83
⋮	⋮	⋮	⋮	⋮

Analysis of the top 20 papers



Future work

- Get bigger data to be able to:
  - Study the parameter dependence beyond λ (in particular, fractional exponent values)
  - Understand the formation of communities (islands?) of highly-valued authors
  - Study and try avoid “undesired consequences”
  - Study the robustness of results (leave one paper out, etc.)

Part 2

The trouble with ad hoc metrics

(the road to hell is paved with good intentions)

Part 2

The trouble with ad hoc metrics
(the road to hell is paved with good intentions)

One example for all: PageRank

- PageRank is essentially a node centrality (importance) measure
- As opposed to node degree, PageRank gives higher weight to links from important nodes (important according to PageRank)

One example for all: PageRank

- PageRank is essentially a node centrality (importance) measure
- As opposed to node degree, PageRank gives higher weight to links from important nodes (important according to PageRank)
- Assign score $p_i^{(t)}$ to each node which initially is uniform: $p_i^{(0)} = 1/N$

$$p_i^{(t+1)} = c \sum_{j \to i} \frac{p_j^{(t)}}{k_j} + \frac{1-c}{N}$$

  - $j \to i$: summation over all nodes j that point to i
  - Here N is the number of nodes and k_j is degree of node j
  - c is a so-called teleportation parameter ($c = 1$: no teleportation)
  - Iterations: convergence is quick even for Google-size networks

Two forms of aging in information networks

- The decay of relevance: $D_R(t)$
  - Node relevance influences the in-coming links
  - Medo et al, PRL 107, 238701 (2011)
  - Medo, PRE 89, 032801 (2014)

Two forms of aging in information networks

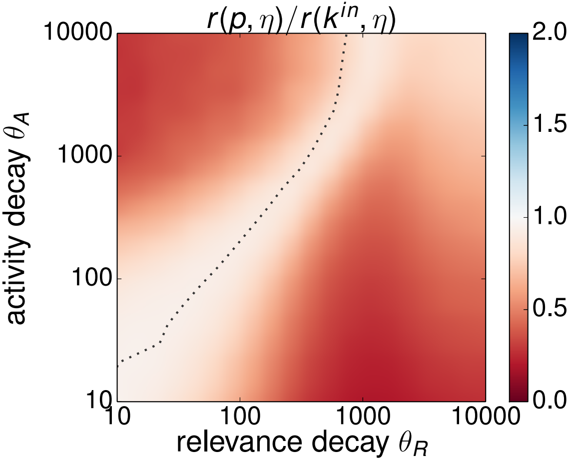
- The decay of relevance: $D_R(t)$
  - Node relevance influences the in-coming links
  - Medo et al, PRL 107, 238701 (2011)
  - Medo, PRE 89, 032801 (2014)
- The decay of activity: $D_A(t)$
  - Nodes activity influences the out-going links

Two forms of aging in information networks

- The decay of relevance: $D_R(t)$
  - Node relevance influences the in-coming links
  - Medo et al, PRL 107, 238701 (2011)
  - Medo, PRE 89, 032801 (2014)
- The decay of activity: $D_A(t)$
  - Nodes activity influences the out-going links
- Assume for simplicity $D_R(t) \sim \exp(-t/\theta_R)$ and $D_A(t) \sim \exp(-t/\theta_A)$
- In the model, each node has intrinsic fitness
- The key question:
  - Can PageRank uncover node fitness?
  - More precisely: Can it do it better than node degree?
  - Practically: Compare $r(p, \eta)$ and $r(k^{in}, \eta)$

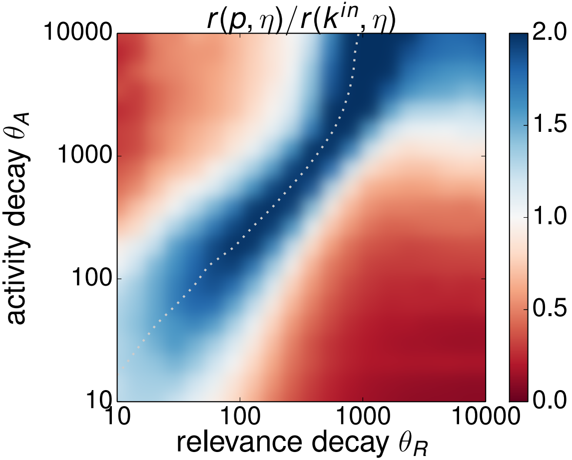
When PageRank fails (Mariani et al, 2016)

PageRank vs indegree in a simple model



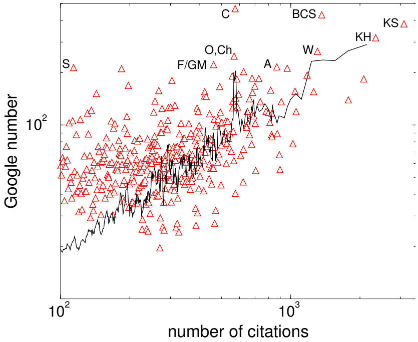
When PageRank fails (Mariani et al, 2016)

PageRank vs indegree in a more complicated model



When PageRank fails: conclusions

1. In citation data, the time scales of relevance and activity decay are very different ($\Theta_A = 0$ because outgoing links are created only upon arrival). PageRank (and its variants) is still commonly applied here…



Chen et al, J Infomet 1, 8 (2007)

When PageRank fails: conclusions

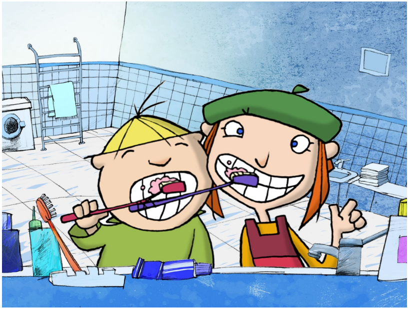
1. In citation data, the time scales of relevance and activity decay are very different ($\Theta_A = 0$ because outgoing links are created only upon arrival). PageRank (and its variants) is still commonly applied here…
2. We need time-dependent metrics/algorithms **based on** and **respecting** microscopical growth rules

When PageRank fails: conclusions

1. In citation data, the time scales of relevance and activity decay are very different ($\Theta_A = 0$ because outgoing links are created only upon arrival). PageRank (and its variants) is still commonly applied here...
2. We need time-dependent metrics/algorithms **based on** and **respecting** microscopical growth rules
3. A lazy solution: Do not compare a paper's PageRank value with values of all other papers but only with papers of similar age

Part 3

Lazy solutions have something about them. . .



From: Lazy Lucy

Correcting PageRank

- Compute PageRank score p for all papers in the APS citation data (1893–2009, 449 937 papers)
- Rescaled PageRank of paper i is

$$\boxed{R_{p,i} = \frac{p_i - \mu_i}{\sigma_i}}$$

  - Here μ_i and σ_i are the mean and standard deviation of p for papers published “close” to paper i
  - Outcome is little sensitive to what “close” means
- Rationale: avoid comparison of apples with oranges

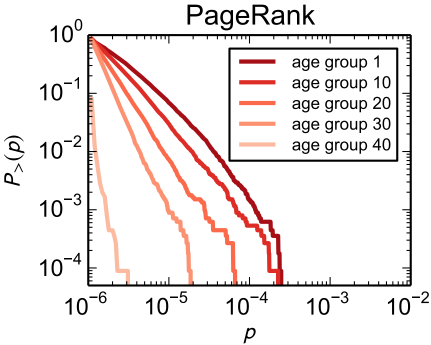
Correcting PageRank

- Compute PageRank score p for all papers in the APS citation data (1893–2009, 449 937 papers)
- Rescaled PageRank of paper i is

$$\boxed{R_{p,i} = \frac{p_i - \mu_i}{\sigma_i}}$$

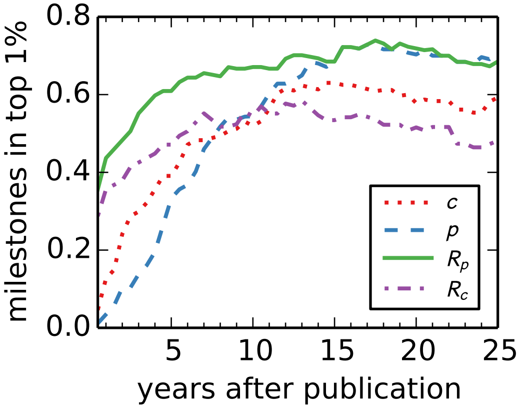
  - Here μ_i and σ_i are the mean and standard deviation of p for papers published "close" to paper i
  - Outcome is little sensitive to what "close" means
- Rationale: avoid comparison of apples with oranges
- Evaluation based on "milestone letters" announced recently (`http://journals.aps.org/prl/50years/milestones`)

Rescaled PageRank: results



Allows us to fairly compare all papers!

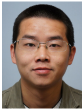
Rescaled PageRank: results



Note: CiteRank is competitive with R_p in some aspects

Thank you for your attention

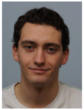
1. H. Liao, R. Xiao, G. Cimini, M. Medo, Network-Driven Reputation in Online Scientific Communities, PLoS ONE 9, e112022, 2014
2. M. Medo, G. Cimini, S. Gualdi, Temporal effects in the growth of networks, Physical Review Letters 107, 238701, 2011
3. M. Medo, Network-based information filtering algorithms: ranking and recommendation, In ”Dynamics on and of Complex Networks 2” (Springer, 2013)
4. M. Medo, Statistical validation of high-dimensional models of growing networks, Physical Review E 89, 032801, 2014
5. M. S. Mariani, M. Medo, Y.-C. Zhang, Ranking nodes in growing networks: When PageRank fails, Scientific Reports 5, 16181, 2015
6. M. S. Mariani, M. Medo, Y.-C. Zhang, Quantifying the significance of scientific papers by time-balanced network centrality (almost submitted)


Hao Liao


Rui Xiao


Giulio Cimini

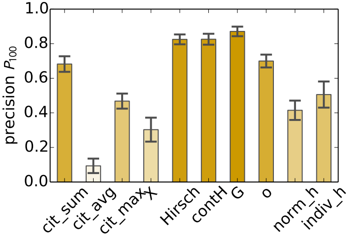

Stanislao Gualdi


Manuel Mariani


Yi-Cheng Zhang

Evaluating researcher performance metrics on artificial datasets (new project with G. Cimini)

- We have good models of information networks
  - Many properties of real datasets can be easily reproduced
- We can use them to grow artificial data of researcher activity
- Goal: compare true researcher "ability" with results of various researcher metrics
- Preliminary results:



- Broader goal: Establish a general simulation and evaluation framework for research activity data