Models and algorithms for growing information networks

Matúš Medo

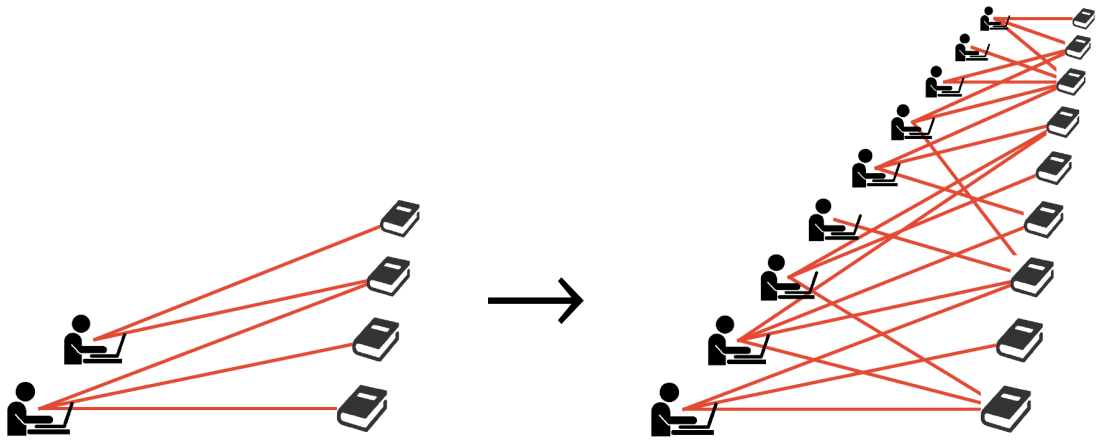
University of Fribourg, Switzerland

The PIIK: 1st Symposium on Network Science

23 November 2016, Zurich

Information networks around us

- E-commerce systems: users and purchased items
- The World Wide Web: web pages
- Citations among scientific papers
- ...



Modeling information networks

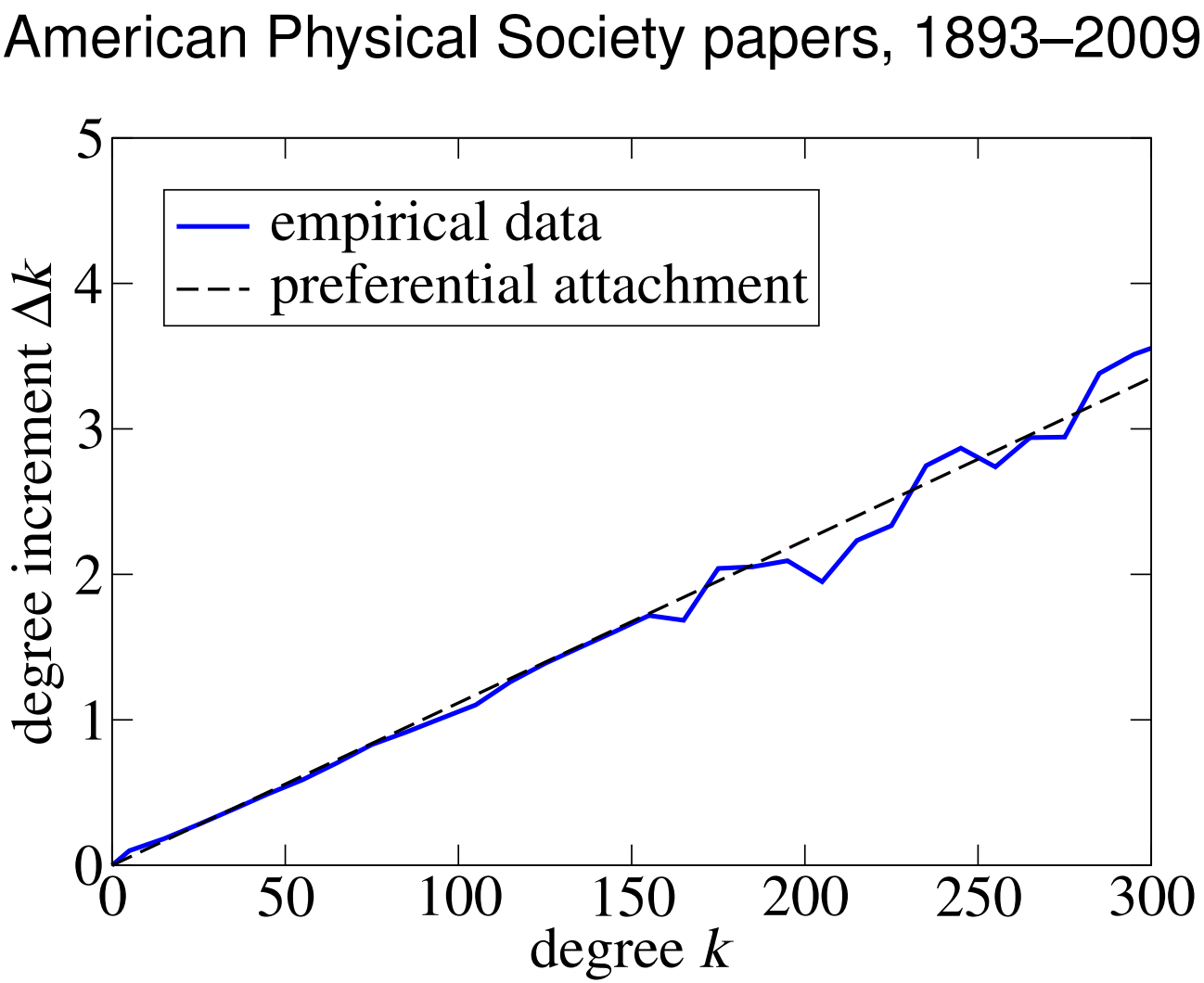
- Preferential attachment
  - Yule (1925), Simon (1955), Price (1976), Barabási & Albert (1999)

Modeling information networks

- Preferential attachment
  - Yule (1925), Simon (1955), Price (1976), Barabási & Albert (1999)
- Probability that item i attracts a new link at time t:

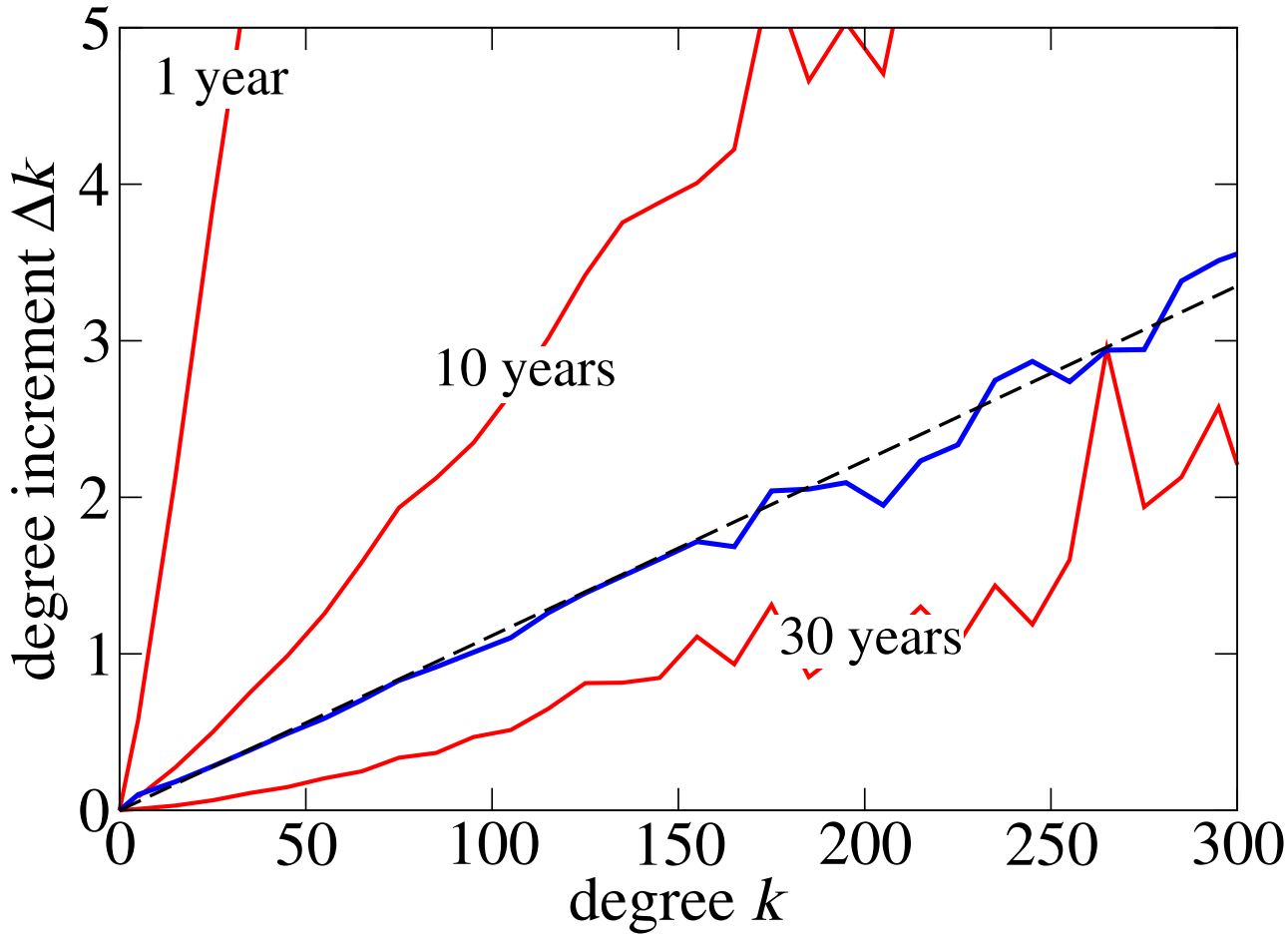
$$P(i,t) \sim \underbrace{k_i(t)}_{\substack{\text{item} \\ \text{degree}}}$$

The missing element



The missing element

American Physical Society papers, 1893–2009



The missing element

"All the News That's Fit to Print."

The New York Times.

THE WEATHER.

TITANIC SINKS FOUR HOURS AFTER HITTING ICEBERG;
866 RESCUED BY CARPATHIA, PROBABLY 1250 PERISH;
ISMAY SAFE, MRS. ASTOR MAYBE, NOTED NAMES MISSING

Col. Astor and Bride, Isidor Straus and Wife, and Maj. Butt Aboard.

"RULE OF SEA" FOLLOWED

PICKED UP AFTER 6 HOURS

Biggest Liner Plunges to the Bottom at 2:20 A.M.

RESCUERS THERE TOO LATE

WOMEN AND CHILDREN FIRST

SEA SEARCH FOR OTHERS

Aging is fundamental

A better model (PRL 107, 238701, 2011)

- Probability that node i attracts a new link

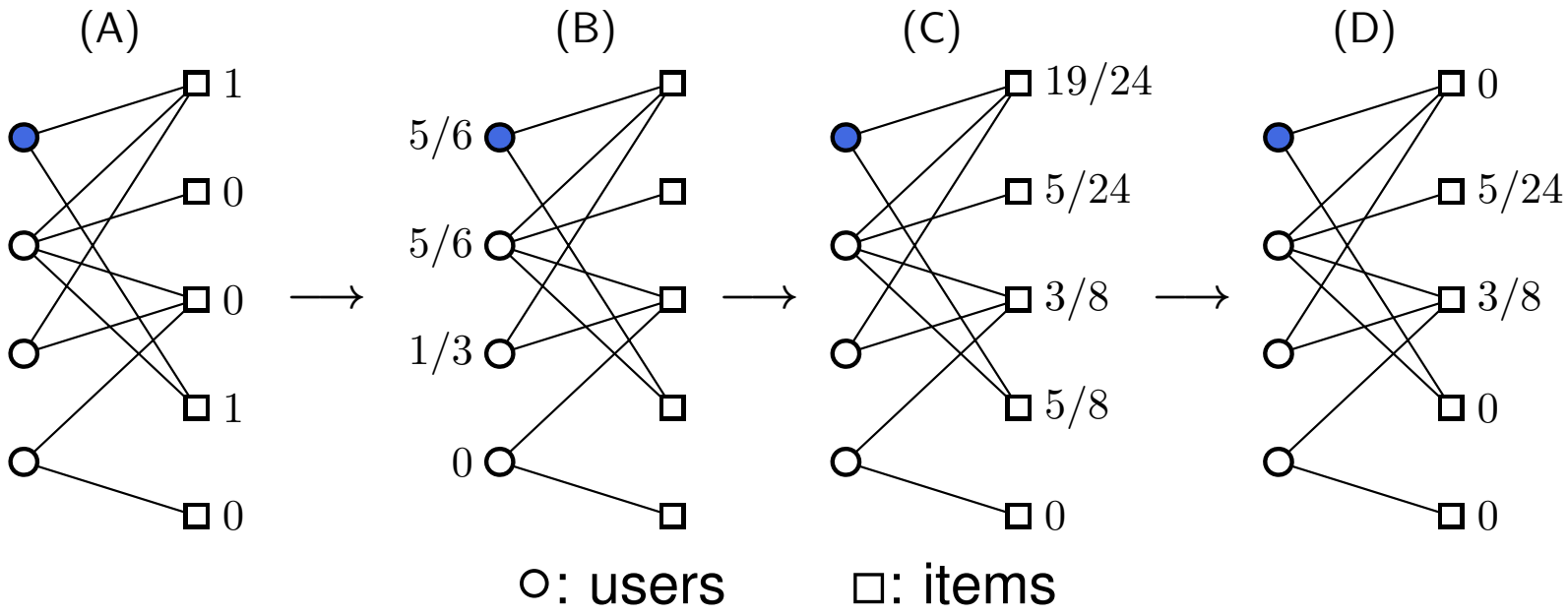
$$P(i,t) \sim \underbrace{k_i(t)}_{\text{degree}} \times \underbrace{D_R(t)}_{\text{aging}} \times \underbrace{f_i}_{\text{fitness}}$$

- The bottom line:
  - Produces realistic degree distributions (power-law, log-normal, etc.)
  - Explains the data better than other models (PRE 89, 032801, 2014)
  - It still does not capture all effects, of course

Application 1: Recommendation

Network-based recommendation

Random walk on a user-item bipartite network



PRE 76, 046115, 2007; PNAS 107, 4511, 2010; Physica A 452, 192, 2016

Two problems with network-based recommendation

1. Ignores time in the recommendation process
2. Ignores time in the evaluation process

Two problems with network-based recommendation

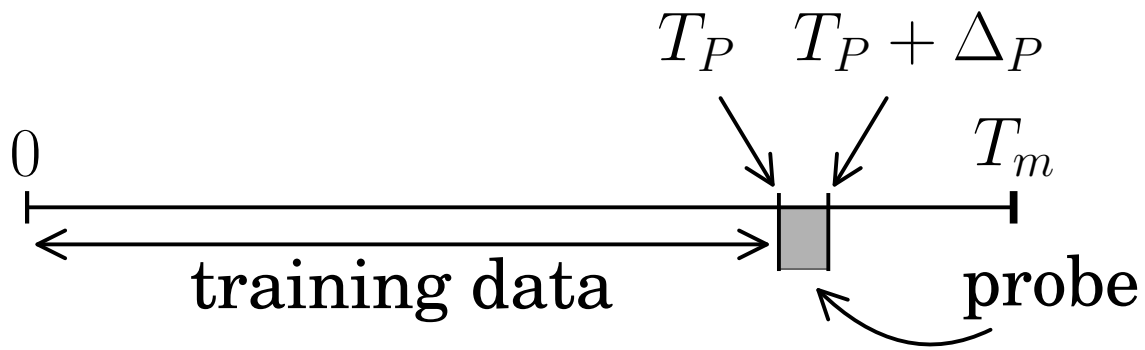
1. Ignores time in the recommendation process
2. Ignores time in the evaluation process

$$\underbrace{u_\alpha^{(i)}}_{\text{new score}} = \underbrace{h_\alpha^{(i)}}_{\text{old score}} \times \underbrace{\frac{\Delta k_\alpha(t,\tau)}{k_\alpha(t)}}_{\text{recent degree increase}}$$

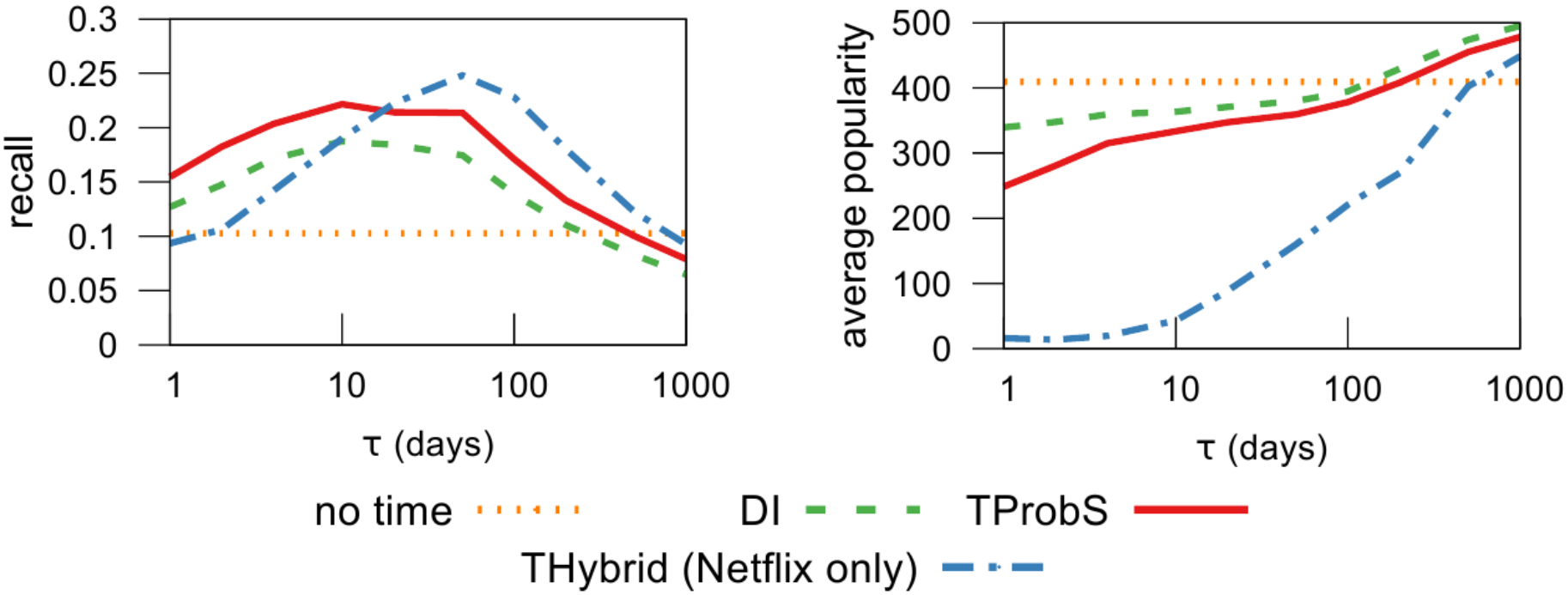
Two problems with network-based recommendation

1. Ignores time in the recommendation process
2. Ignores time in the evaluation process

$$\underbrace{u_\alpha^{(i)}}_{\text{new score}} = \underbrace{h_\alpha^{(i)}}_{\text{old score}} \times \underbrace{\frac{\Delta k_\alpha(t,\tau)}{k_\alpha(t)}}_{\text{recent degree increase}}$$



Two-fold improvement (to appear in EPL)



τ is the size of the time window to compute the temporal features

Application 2:
Ranking network nodes

PageRank: A classical network centrality metric

- Centrality metrics quantify the importance of nodes
- Simplest centrality metric: in-degree
- PageRank gives higher weight to links from important nodes

PageRank: A classical network centrality metric

- Centrality metrics quantify the importance of nodes
- Simplest centrality metric: in-degree
- PageRank gives higher weight to links from important nodes

- PageRank score p_i of node i is

$$p_i = \underbrace{c \sum_{j \to i} \frac{p_j}{k_j}}_{\text{network contribution}} + \underbrace{\frac{1-c}{N}}_{\textit{teleportation}}$$

  - $c = 0.85$ (WWW) or $c = 0.5$ (citation networks)
  - Solvable even for Google-size networks

Evaluation on model networks

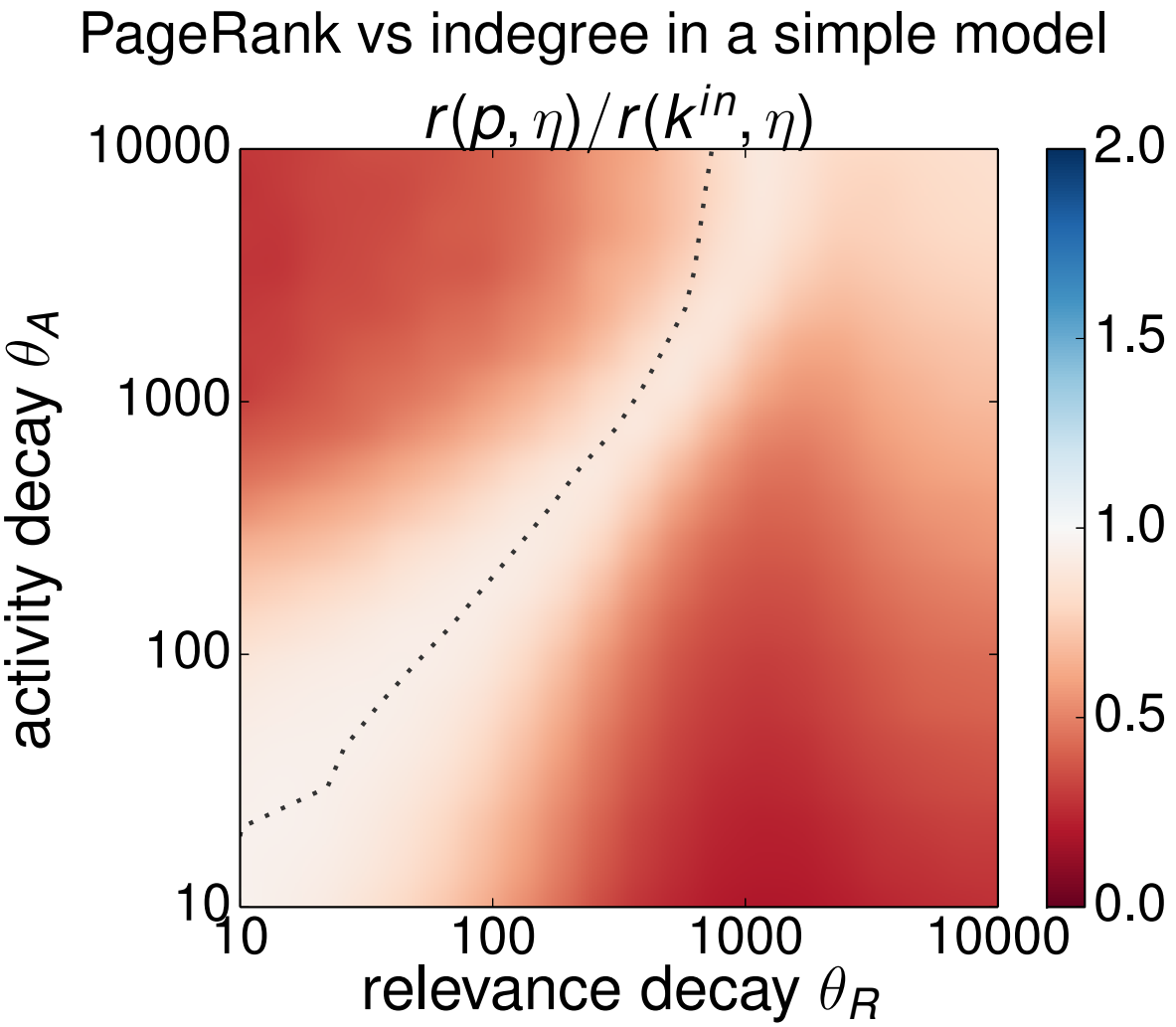
- Three key elements of the model:
  1. Node i has intrinsic fitness η_i (before f_i)
  2. Decay of relevance (attractiveness to incoming links): $D_R(t)$
  3. Decay of activity (activity to create outgoing links): $D_A(t)$
- We assume $D_R(t) \sim \exp(-t/\theta_R)$ and $D_A(t) \sim \exp(-t/\theta_A)$

Evaluation on model networks

- Three key elements of the model:
  1. Node i has intrinsic fitness η_i (before f_i)
  2. Decay of relevance (attractiveness to incoming links): $D_R(t)$
  3. Decay of activity (activity to create outgoing links): $D_A(t)$
- We assume $D_R(t) \sim \exp(-t/\theta_R)$ and $D_A(t) \sim \exp(-t/\theta_A)$

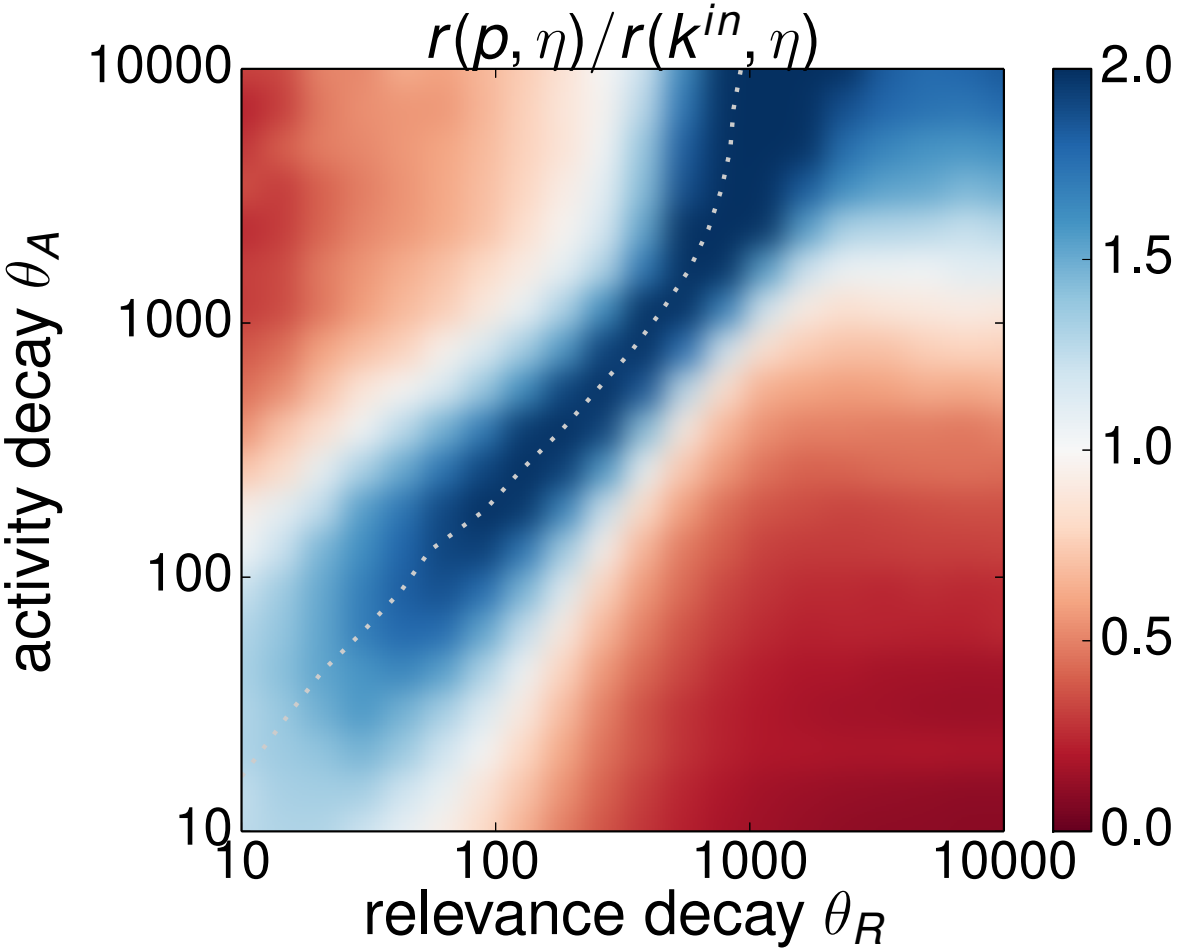
- The key question: Can PageRank uncover node fitness?
  - More precisely: Can it do it better than node degree?
  - Practically: Evaluate $r(p, \eta)$ / $r(k^{in}, \eta)$

When PageRank fails (Scientific Reports, 2016)



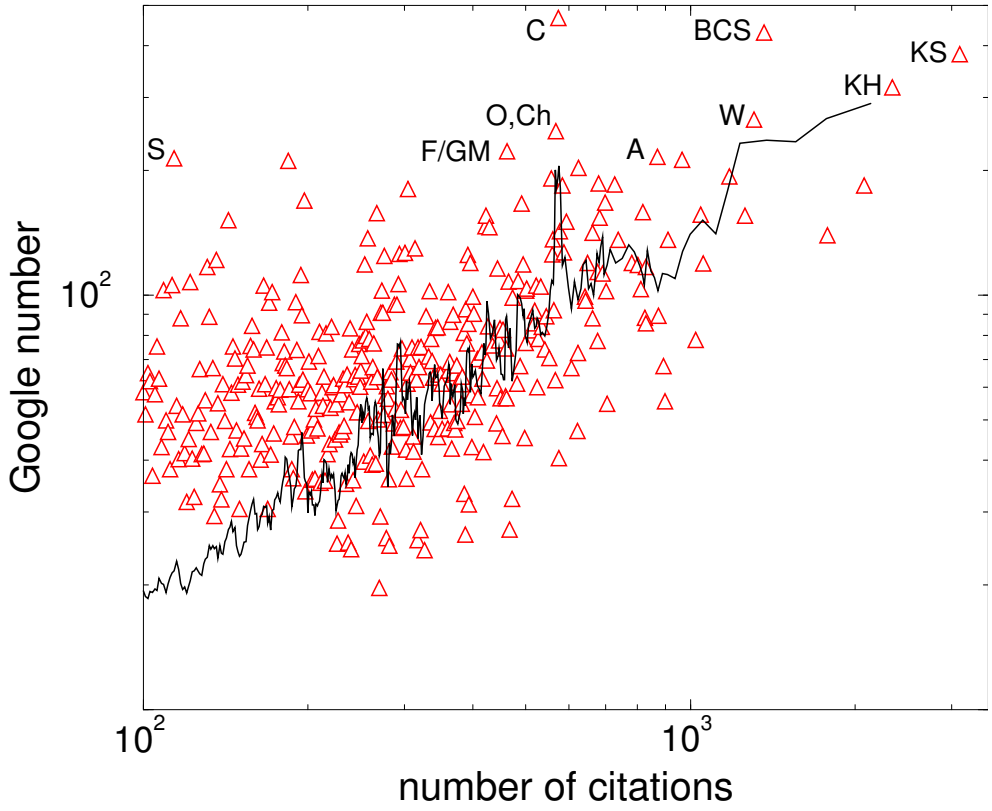
When PageRank fails (Scientific Reports, 2016)

PageRank vs indegree in a more complicated model



When PageRank fails: conclusions

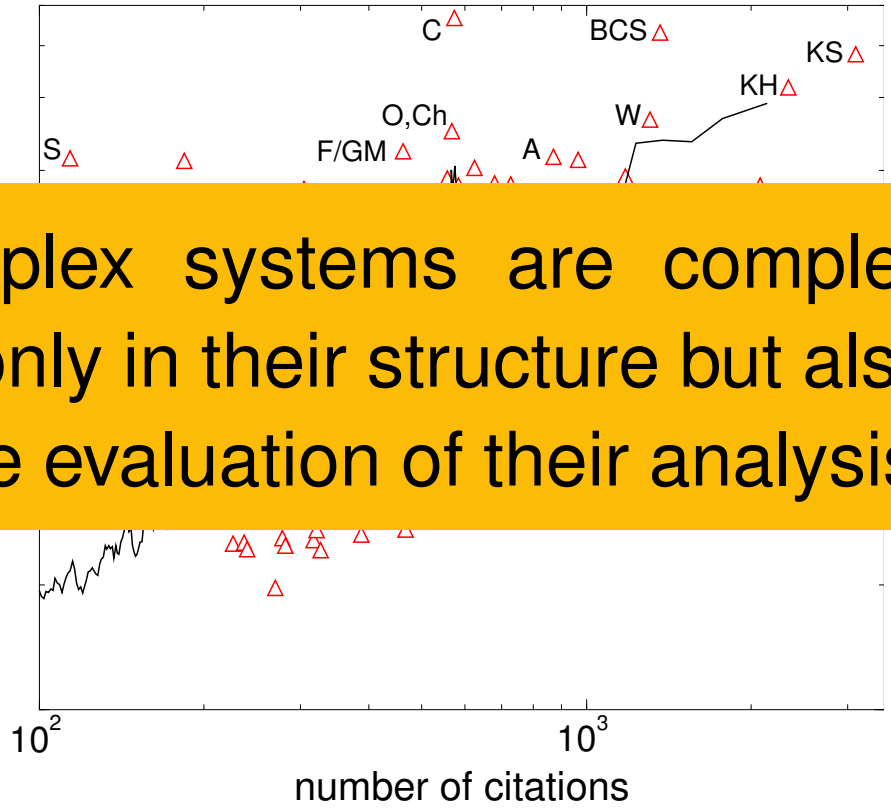
1. Citation data fall in a very wrong part of the (Θ_R, Θ_A) plane, yet PageRank is still commonly applied there...



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

When PageRank fails: conclusions

1. Citation data fall in a very wrong part of the (Θ_R, Θ_A) plane, yet PageRank is still commonly applied there...



Complex systems are complex not only in their structure but also in the evaluation of their analysis.

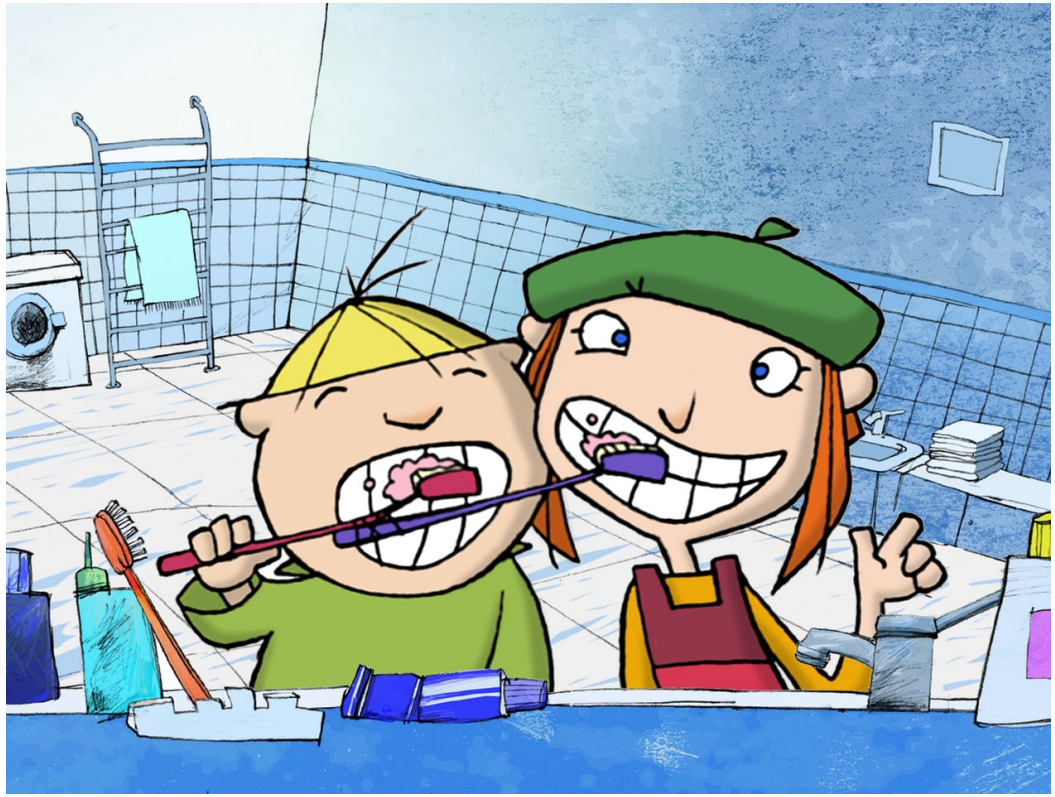
Chen et al, J Infomet 1, 8 (2007)

When PageRank fails: conclusions

1. Citation data fall in a very wrong part of the (Θ_R, Θ_A) plane, yet PageRank is still commonly applied there...
2. We need time-dependent metrics/algorithms *based on* and *respecting* the microscopical growth rules

When PageRank fails: conclusions

1. Citation data fall in a very wrong part of the (Θ_R, Θ_A) plane, yet PageRank is still commonly applied there...
2. We need time-dependent metrics/algorithms *based on* and *respecting* the 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



From: Lazy Lucy

Let’s be lazy for once. . .

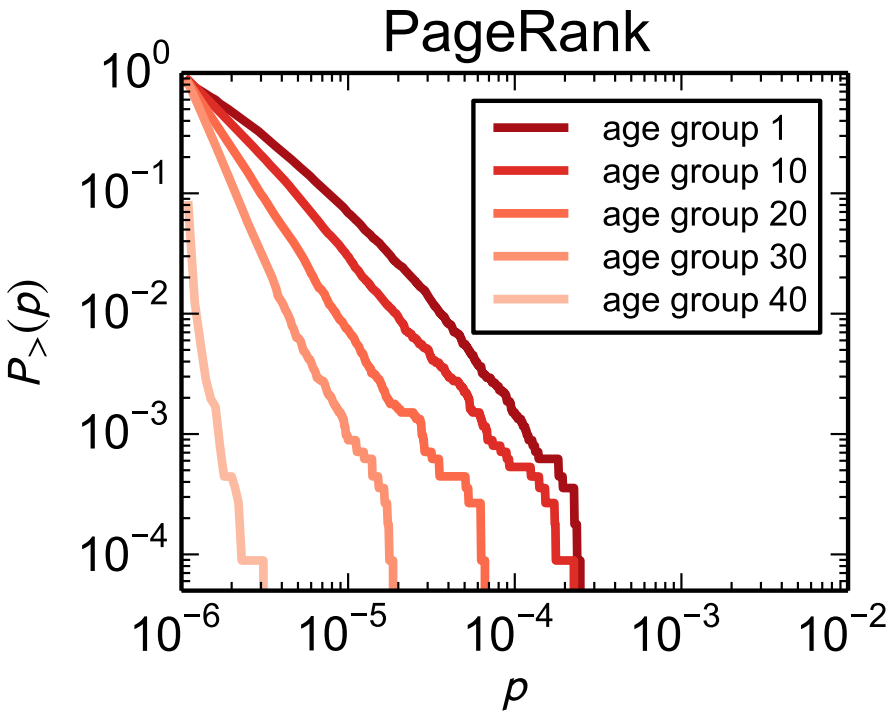
Correcting PageRank (Journal of Informetrics, 2016)

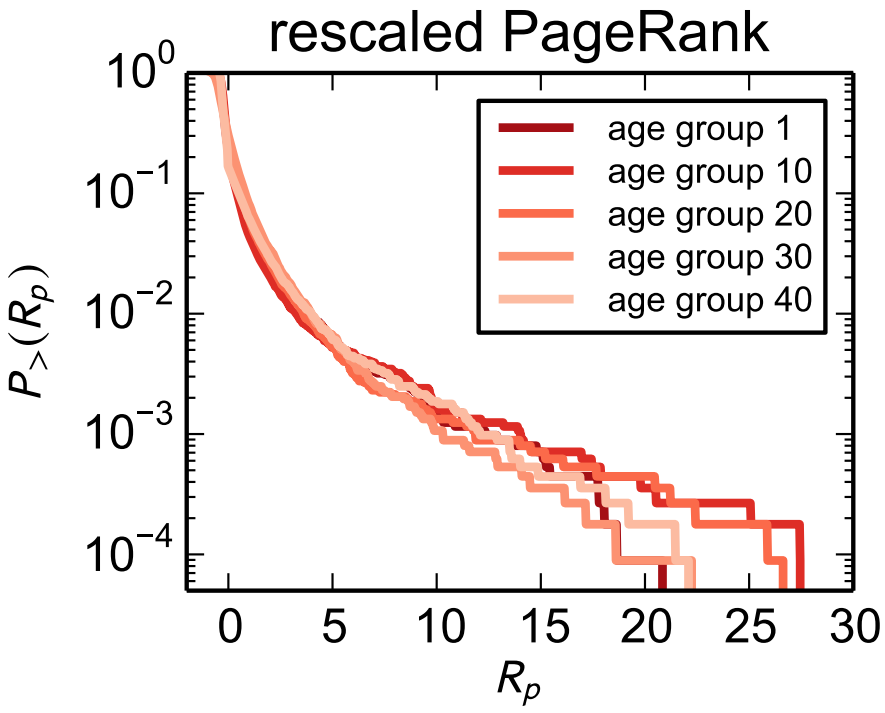
- Compute PageRank score p for all papers in the APS citation data (1893–2009, 449 937 papers)
- Rescaled PageRank of paper i is

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

  - 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
  - Our close: a window of 1000 papers around i
- Rationale: avoid comparison of apples with oranges

Rescaled PageRank: bias removal



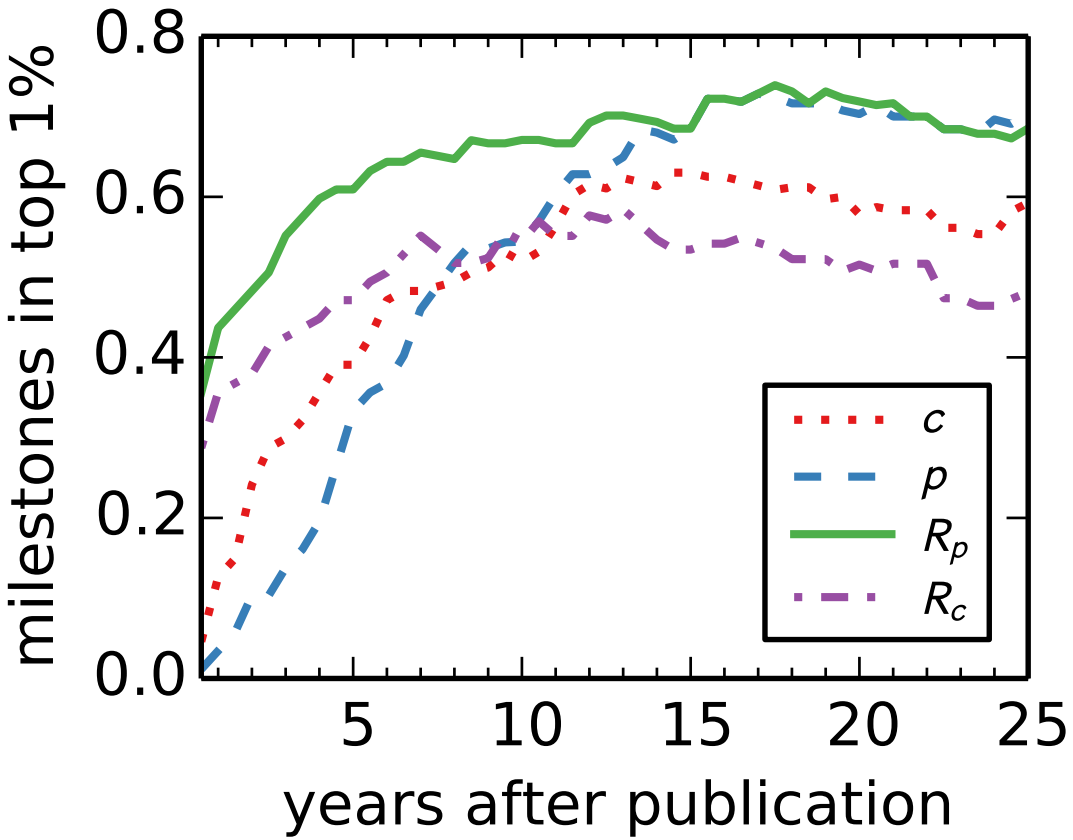


↙

Allows us to fairly compare all papers!

Rescaled PageRank: identification of milestones

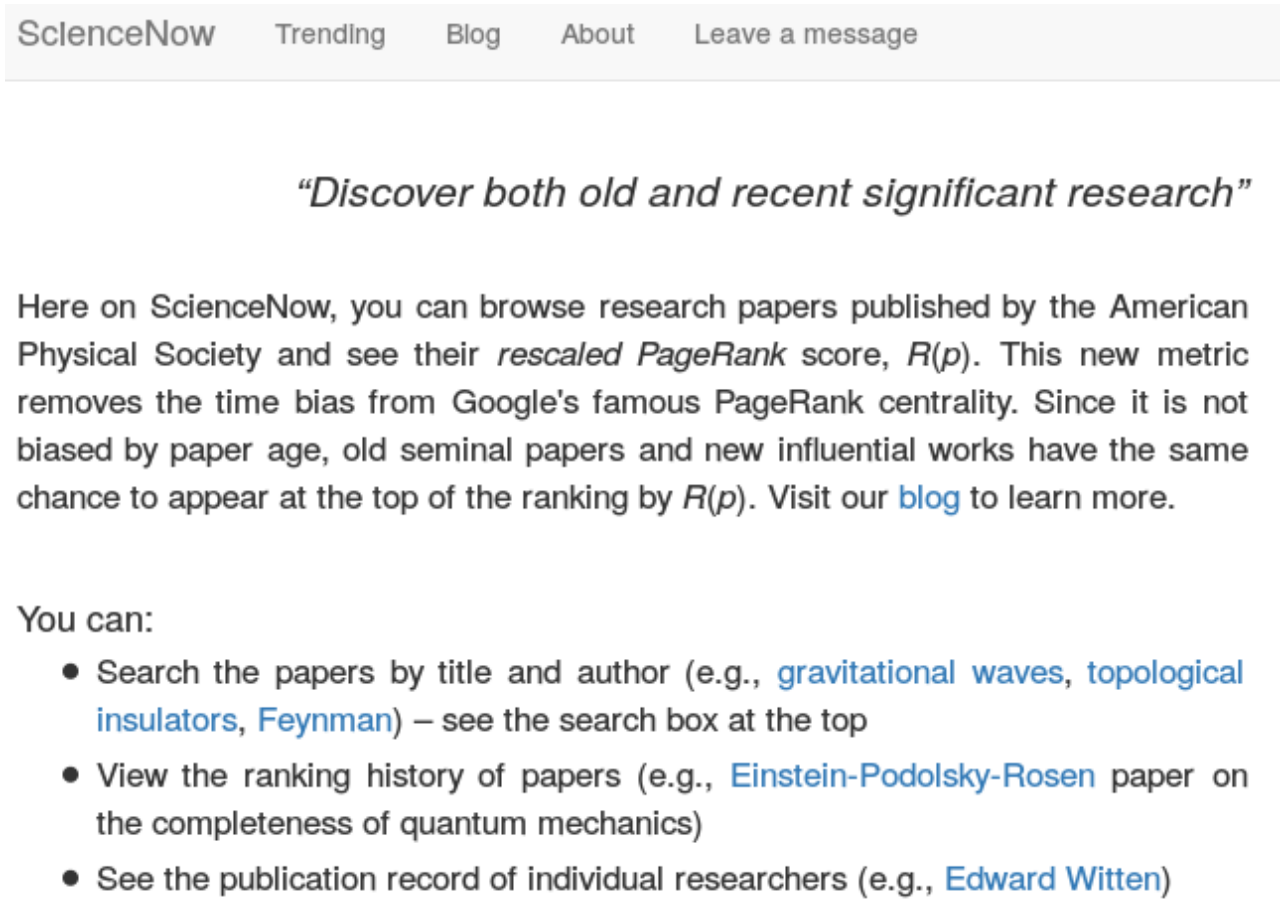
Evaluation based on a list of “milestone letters” announced by PRL



Note: CiteRank is competitive with R_p in some aspects

From research to applications

Browse 600,000 physics papers at www.sciencenow.info



ScienceNow Trending Blog About Leave a message

"Discover both old and recent significant research"

Here on ScienceNow, you can browse research papers published by the American Physical Society and see their *rescaled PageRank* score, $R(p)$. This new metric removes the time bias from Google's famous PageRank centrality. Since it is not biased by paper age, old seminal papers and new influential works have the same chance to appear at the top of the ranking by $R(p)$. Visit our blog to learn more.

You can:

- Search the papers by title and author (e.g., gravitational waves, topological insulators, Feynman) – see the search box at the top
- View the ranking history of papers (e.g., Einstein-Podolsky-Rosen paper on the completeness of quantum mechanics)
- See the publication record of individual researchers (e.g., Edward Witten)

Three take-away points

1. Time dimension is fundamental in information networks
2. Beware the application range of “good old” metrics
3. By including time, we can do better

Complex systems
as static objects

→

Complex systems
as evolving objects

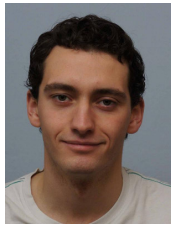
Further related works:

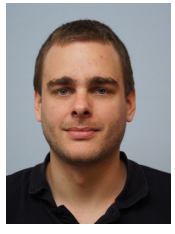
1. M. Medo, G. Cimini, Model-based evaluation of scientific impact indicators, Physical Review E 94, 032312, 2016
2. M. Medo *et al.*, Identification and modeling of discoverers in online social systems, Scientific Reports 6, 34218, 2016
3. A. Vidmer *et al.*, Unbiased metrics of friends' influence in multi-level networks, EPJ Data Science 4, 20, 2015

Web site: www.ddp.fmph.uniba.sk/~medo/physics/


Giulio Cimini


Stanislao Gualdi


Manuel Mariani


Alex Vidmer


An Zeng


Yi-Cheng Zhang

Thank you for your attention!