Information networks: from data to models and algorithms

Matúš Medo

University of Fribourg, Switzerland

Challenges in Data Science: A complex systems perspective

14-17 October 2015, Torino

Outline

1. Growing networks with fitness and aging
2. Temporal bias of PageRank
3. Discoveries and discoverers in social systems

Outline

1. Growing networks with fitness and aging
2. Temporal bias of PageRank
3. Discoveries and discoverers in social systems

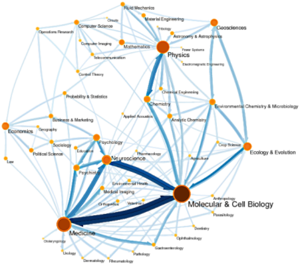
The common theme

Temporal patterns in information and social systems.

We live the information age

Part 1

Growth of information networks



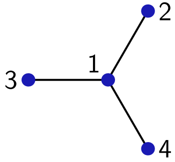
Preferential attachment (PA)

- A classical network model
  - Yule (1925), Simon (1955), Price (1976), Barabási & Albert (1999)
  - Growth of cities, citations of scientific papers, WWW,...

Preferential attachment (PA)

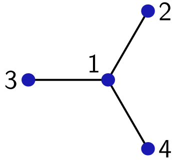
- A classical network model
  - Yule (1925), Simon (1955), Price (1976), Barabási & Albert (1999)
  - Growth of cities, citations of scientific papers, WWW,...
- Nodes and links are added with time
- Probability that a node acquires a new link proportional to its current degree

$$P(i,t) \sim k_i(t)$$



Preferential attachment (PA)

- A classical network model
  - Yule (1925), Simon (1955), Price (1976), Barabási & Albert (1999)
  - Growth of cities, citations of scientific papers, WWW,...
- Nodes and links are added with time
- Probability that a node acquires a new link proportional to its current degree

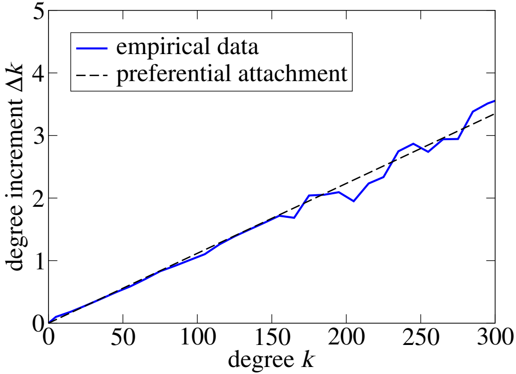


$$P(i, t) \sim k_i(t)$$

- Pros: simple, produces a power-law degree distribution
- Cons: The power-law degree distribution due to the first nodes

PA in scientific citation data

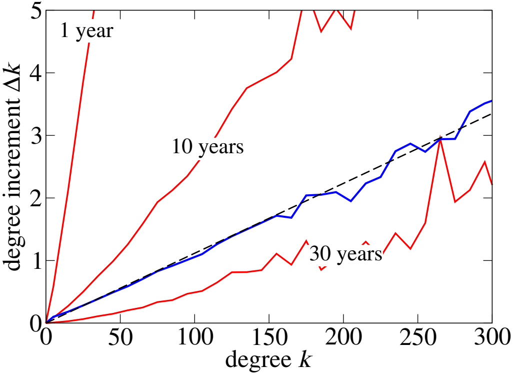
Journals of the American Physical Society from 1893 to 2009:



See also Adamic & Huberman (2000), Redner (2005), Newman (2009),. . .

PA in scientific citation data

Journals of the American Physical Society from 1893 to 2009:



Time decay is fundamental


"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

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

RESCUERS THERE TOO LATE

WOMEN AND CHILDREN FIRST

Growing networks with fitness and aging (PRL 107, 238701, 2011)

- Probability that node i attracts a new link

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

- The aging factor $D_R(t)$ decays with time: a decay of relevance

Growing networks with fitness and aging (PRL 107, 238701, 2011)

- Probability that node *i* attracts a new link

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

  - The aging factor $D_R(t)$ decays with time: a decay of relevance

- The bottom line:
  - **Good:** Produces various realistic degree distributions (power-law, etc.)
  - **Bad:** Difficult to validate (high-dimensional statistics)
  - **Good:** This model explains the data much better than any other (Medo, Phys Rev E 89, 032801, 2014)

Fitness and aging: conclusions

- PA with fitness and aging as a relevant model for *information* networks
- There are many possible applications: ranking, prediction, ...
- Even better: this establishes a playground!

Part 2

Temporal bias of PageRank

What is PageRank

- PageRank is essentially a node centrality (importance) measure
- Simplest centrality: degree (counting the links—local)

What is PageRank

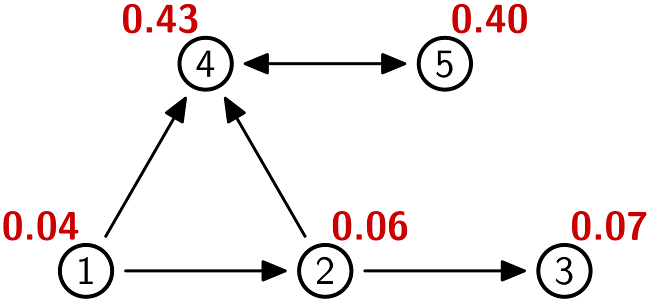
- PageRank is essentially a node centrality (importance) measure
- Simplest centrality: degree (counting the links—local)
- Non-local: PageRank (links from important nodes count more)
- Assign score $p_i^{(t)}$ to each node which initially is uniform: $p_i^{(0)} = 1/N$

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

  - $j \to i$ are 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 quick even for Google-size networks

What is PageRank

- PageRank is essentially a node centrality (importance) measure
- Simplest centrality: degree (counting the links—local)



Important nodes are those that are pointed by other important nodes

Two forms of aging in information networks

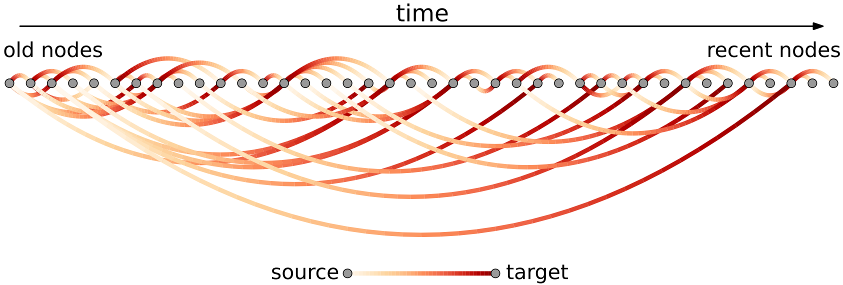
- The decay of relevance: $D_R(t)$
  - Node relevance influences the in-coming links

Two forms of aging in information networks

- The decay of relevance: $D_R(t)$
  - Node relevance influences the in-coming links
- 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
- The decay of activity: $D_A(t)$
  - Nodes activity influences the out-going links



A growing network with a quick decay of attractiveness and no decay of activity

A model to test the effect of aging

- In both cases, we assign fitness f_i and activity A_i to nodes
- Aging applies to both: $D_R(t) = \exp(-t/\theta_R)$ and $D_A(t) = \exp(-t/\theta_A)$

A model to test the effect of aging

- In both cases, we assign fitness f_i and activity A_i to nodes
- Aging applies to both: $D_R(t) = \exp(-t/\theta_R)$ and $D_A(t) = \exp(-t/\theta_A)$
- The probability of node i to create an outgoing link is

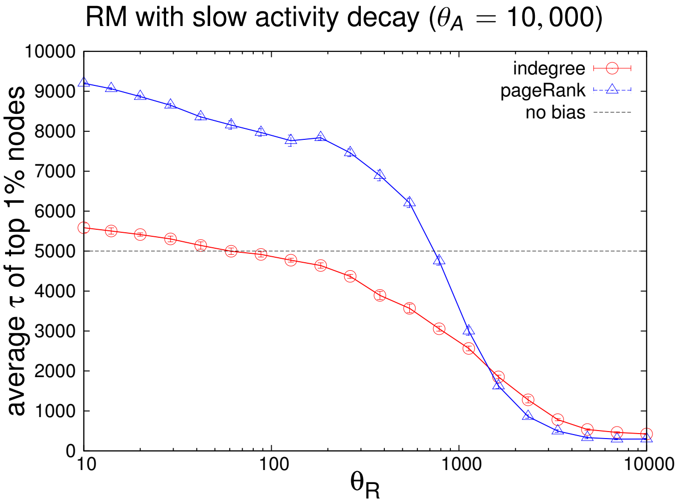
$$P_i^{out} \sim A_i D_A(t - \tau_i)$$

- The probability of node j to receive an incoming link is

$$P_j^{in}(t) \sim (k_j^{in}(t) + 1)\, f_j\, D_R(t - \tau_j)$$

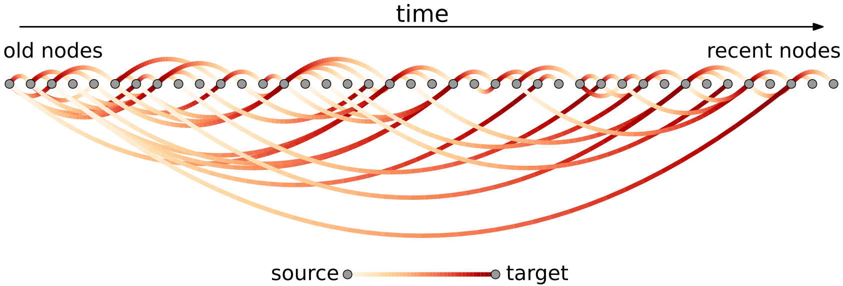
  - This is our old friend: the relevance model (RM)
  - A small modification of RM, extended fitness model (EFM), is more suitable for PageRank use

The biases of PageRank

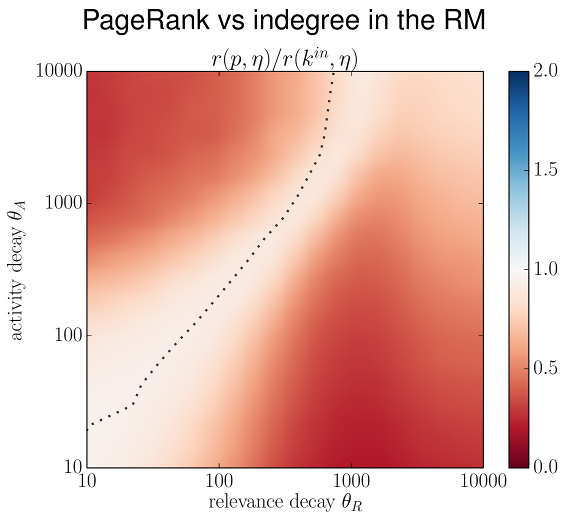


The biases of PageRank

Why the new kind of bias?

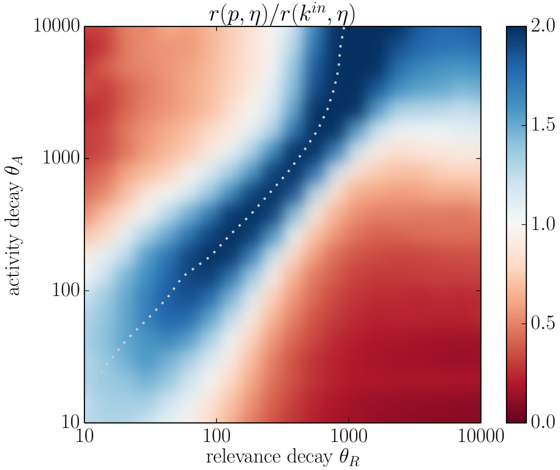


The biases of PageRank



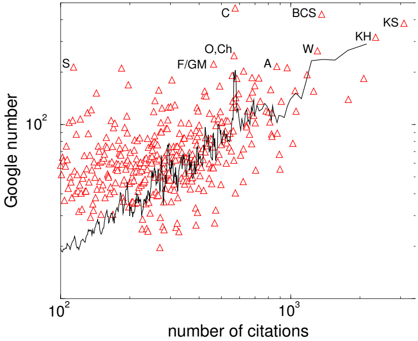
The biases of PageRank

PageRank vs indegree in the EFM



The biases of PageRank: 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)

The biases of PageRank: 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 algorithms based on microscopical growth rules

The biases of PageRank: 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 algorithms based on 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. Preliminary results seem very promising (see the poster)...

Part 3

Discoverers in online social systems

Beyond preferential attachment in social systems

- Bipartite user-item data (e.g., *who* bought *what* at Amazon.com)
  - Similar behavior in monopartite social data (user-user)
- Previous research shows/assumes that users are driven by popularity combined with fitness and aging

Beyond preferential attachment in social systems

- Bipartite user-item data (e.g., *who* bought *what* at Amazon.com)
  - Similar behavior in monopartite social data (user-user)
- Previous research shows/assumes that users are driven by popularity combined with fitness and aging

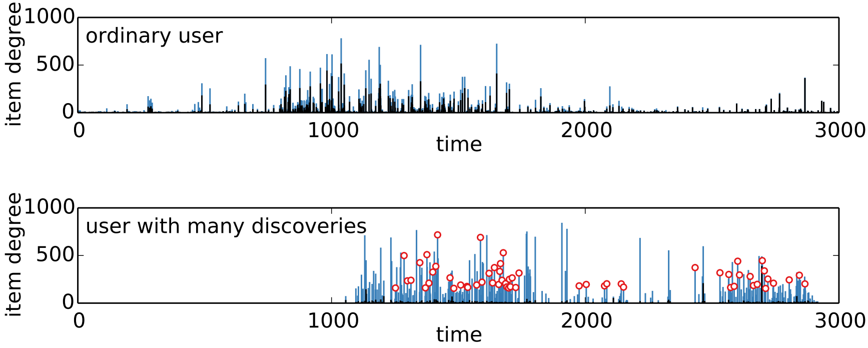
But is this the whole story?

Beyond preferential attachment in social systems

- Bipartite user-item data (e.g., *who* bought *what* at Amazon.com)
  - Similar behavior in monopartite social data (user-user)
- Previous research shows/assumes that users are driven by popularity combined with fitness and aging

- To find the users who defy popularity, we do the following:
  - A user makes a *discovery* when they are among the first 5 users to collect an eventually highly popular item (top 1% of all items are used as target)
  - A new metric, *user surprisal*, shows that there are users who make discoveries so often that it cannot be explained by luck

Discoveries in Amazon data



Black bars: popularity of collected items when they are collected.
Blue bars: final popularity of collected items.
Red circles: discoveries.

How to quantify the user success

- This concept yields the *number of discoveries* d_i for each user
- We also know the *number of links* k_i made by each user
- How to assess how unusual is a given user?

How to quantify the user success

- This concept yields the *number of discoveries d_i* for each user
- We also know the *number of links k_i* made by each user
- How to assess how unusual is a given user?
- The overall discovery probability is $p_D = D/L$
  - Here $D = \sum_i d_i$ and $L = \sum_i k_i$
- Assuming that all users and links are equal, the probability that a user makes *at least d_i discoveries in k_i attempts* is

$$P^0(d_i|k_i, p_D, H_0) = \sum_{n=d_i}^{k_i} \binom{k_i}{n} p_D^n (1 - p_D)^{k_i - n}$$

- Motivated by information theory, we introduce user surprisal

$$s_i := -\ln P^0(d_i|k_i, p_D, H_0)$$

Top users in the Amazon data

Rank	k_i	d_i	r_i	P_i^0	s_i
1	188	59	51.6	10^{-82}	187.6
2	244	50	33.7	10^{-59}	135.3
3	217	35	26.5	10^{-38}	86.4
4	237	26	18.0	10^{-24}	54.4
5	190	24	20.8	10^{-24}	53.8
6	364	26	11.7	10^{-19}	43.5
7	185	18	16.0	10^{-16}	36.1
8	73	11	24.8	10^{-12}	27.6
9	41	9	36.1	10^{-12}	26.4
10	60	10	27.4	10^{-12}	26.2
			...		

Top users in the Amazon data

Rank	k_i	d_i	r_i	P_i^0	s_i
1	188	59	51.6	10^{-82}	187.6
2	244	50	33.7	10^{-59}	135.3
3	217	35	26.5	10^{-38}	86.4
4	237	26	18.0	10^{-24}	54.4
5	190	24	20.8	10^{-24}	53.8
6	364	26	11.7	10^{-19}	43.5
7	185	18	16.0	10^{-16}	36.1
8	73	11	24.8	10^{-12}	27.6
9	41	9	36.1	10^{-12}	26.4
10	60	10	27.4	10^{-12}	26.2
			...		

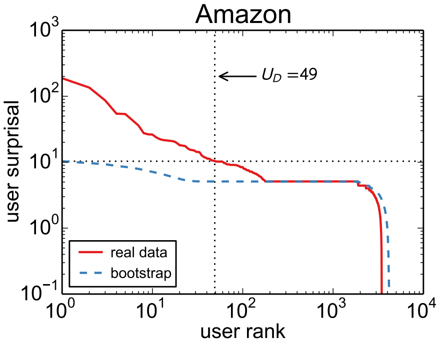
But: Is this not just luck?

Discoverer or a lucky guy?

1. We generate the users' number of discoveries under the null hypothesis
2. The generated data are then used to compute "bootstrap" user surprisal values
3. We check whether a user's real surprisal is higher than the average highest surprisal in bootstrap realizations

Discoverer or a lucky guy?

1. We generate the users' number of discoveries under the null hypothesis
2. The generated data are then used to compute "bootstrap" user surprisal values
3. We check whether a user's real surprisal is higher than the average highest surprisal in bootstrap realizations



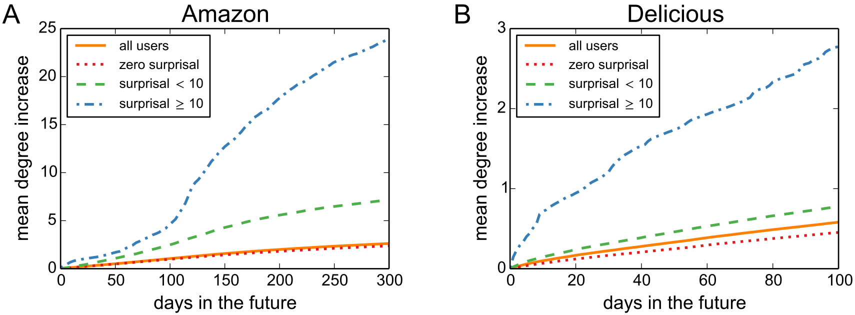
Is this any useful?

Take young items with only one link and divide them into groups depending on the surprisal of the user who has collected them

With such extremely limited information (only the first link for each item), predictions are difficult, especially about the future...

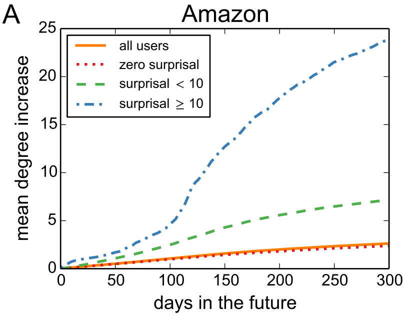
Is this any useful?

Take young items with only one link and divide them into groups depending on the surprisal of the user who has collected them



Is this any useful?

Take young items with only one link and divide them into groups depending on the surprisal of the user who has collected them



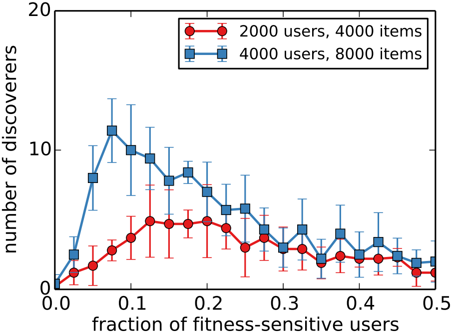
The answer: Yes, potentially very useful!

A network model

- Network growth model with to rules reproduces the real data patterns
  1. Some users are popularity-driven: $k_i(t)D_R(t)$
  2. Others are fitness-driven: $f_i(t)D_R(t)$

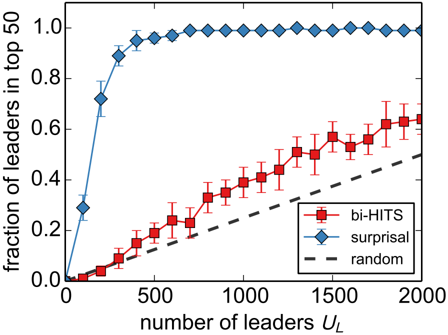
A network model

- Network growth model with to rules reproduces the real data patterns
  1. Some users are popularity-driven: $k_i(t)D_R(t)$
  2. Others are fitness-driven: $f_i(t)D_R(t)$
- The discoverer behavior can be reproduced



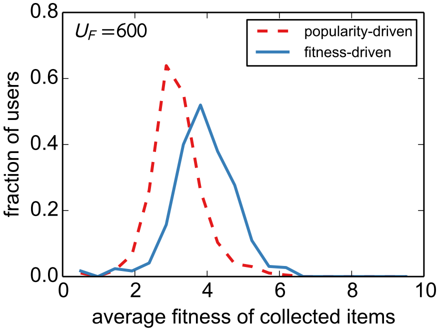
A network model

- Network growth model with to rules reproduces the real data patterns
  1. Some users are popularity-driven: $k_i(t)D_R(t)$
  2. Others are fitness-driven: $f_i(t)D_R(t)$
- Model data pose a puzzle to classical ranking algorithms



A network model

- Network growth model with to rules reproduces the real data patterns
  1. Some users are popularity-driven: $k_i(t)D_R(t)$
  2. Others are fitness-driven: $f_i(t)D_R(t)$
- Model data pose a puzzle to classical ranking algorithms



Reason: Insightful choices of the leaders are copied by the followers. All users ultimately collect items of the same fitness and an algorithm acting on a static data snapshot cannot distinguish them.
Solution: Algorithms that take time into account adequately.

Discoverers: conclusions

- We find discoverers in almost any information network we look at
- There are still many open questions. . .

Discoverers: conclusions

- We find discoverers in almost any information network we look at
- There are still many open questions. . .
  1. What other influences contribute to the observed discovery patterns? Social status? Do the users have head start on some items?
  2. How best to decide who is a discoverer and who is not?
  3. How best to use this information for popularity prediction?
  4. How to model this kind of data?
     E.g., to which extend do the ordinary users ignore fitness?
  5. How does all this translates to monopartite data?
  6. There is fine struture—someone is maybe a discoverer in sci-fi movies but very ordinary in romantic movies; how to approach this?
  7. How to use this knowledge to design better algorithms?

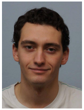
Thank you for your attention

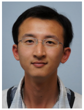
1. M. Medo, G. Cimini, S. Gualdi, Temporal effects in the growth of networks, Physical Review Letters 107, 238701, 2011
2. Y. Berset, M. Medo, The effect of the initial network configuration on preferential attachment, European Physical Journal B 86, 260, 2013
3. M. Medo, Network-based information filtering 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, arXiv:1509.01476 (accepted in Scientific Reports)
6. M. Medo, M. S. Mariani, A. Zeng, Y.-C. Zhang, Identification and modeling of discoverers in online social systems, arXiv:1509.01477


Yves Berset


Giulio Cimini


Stanislao Gualdi


Manuel Mariani


An Zeng


Yi-Cheng Zhang