Growing networks as models for information and social systems

Matúš Medo

University of Fribourg, Switzerland

GROWTHCOM Project Summer School

10 September 2015, Lipari

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 and the role of time in information and social systems.

Outline

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

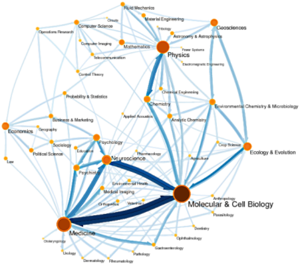
The common theme

Temporal patterns and the role of time in information and social systems.

This is not as much about economics but still: information is vital for business. What we do is very relevant to e-commerce, for example.

Part 1

Growth of information networks



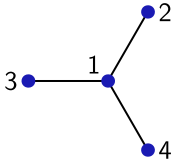
Preferential attachment (PA)

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

Preferential attachment (PA)

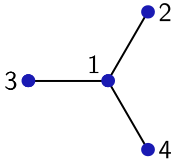
- A classical network model
  - Yule (1925), Simon (1955), Price (1976), Barabási & Albert (1999)
  - Growth of cities, citations of scientific papers, WWW,...
- Nodes and links are added with time
- Probability that a node acquires a new link proportional to its current degree

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



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



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

- In detail: $P(i,t) = \frac{k_i(t)}{\sum_j k_j(t)}$ or $P(i,t) = \frac{k_i(t)+C}{\sum_j [k_j(t)+C]}$
- Pros: simple, produces a power-law degree distribution

Solving the basic PA model: the contiuum approach

1. Starting with two connected nodes at time 1, we introduce one node at each time step and connect it with one existing node

Solving the basic PA model: the contiuum approach

1. Starting with two connected nodes at time 1, we introduce one node at each time step and connect it with one existing node
2. Approximate the probabilistic degree evolution with the average one

$$\frac{\mathrm{d}\overline{k_i(t)}}{\mathrm{d}t} = \frac{\overline{k_i(t)}}{\sum_j k_j(t)} = \frac{\overline{k_i(t)}}{2t} \implies \overline{k_i(t)} \sim \sqrt{t}$$

Solving the basic PA model: the contiuum approach

1. Starting with two connected nodes at time 1, we introduce one node at each time step and connect it with one existing node
2. Approximate the probabilistic degree evolution with the average one

$$\frac{\mathrm{d}\overline{k_i(t)}}{\mathrm{d}t} = \frac{\overline{k_i(t)}}{\sum_j k_j(t)} = \frac{\overline{k_i(t)}}{2t} \implies \overline{k_i(t)} \sim \sqrt{t}$$

3. The initial condition is $k_i(i) = 1$, hence $\overline{k_i(t)} = \sqrt{t/i}$

Solving the basic PA model: the contiuum approach

1. Starting with two connected nodes at time 1, we introduce one node at each time step and connect it with one existing node
2. Approximate the probabilistic degree evolution with the average one

$$\frac{\mathrm{d}\overline{k_i(t)}}{\mathrm{d}t} = \frac{\overline{k_i(t)}}{\sum_j k_j(t)} = \frac{\overline{k_i(t)}}{2t} \implies \overline{k_i(t)} \sim \sqrt{t}$$

3. The initial condition is $k_i(i) = 1$, hence $\overline{k_i(t)} = \sqrt{t/i}$
4. Now the distribution of i is uniform among the nodes

$$P(k)\,\mathrm{d}k = \varrho(i)\,\mathrm{d}i \implies P(k) = \varrho(i)\left|\frac{\mathrm{d}k}{\mathrm{d}i}\right|^{-1} \sim i^{3/2} \sim k^{-3}$$

Solving the basic PA model: the master equation

1. Say $p_{k,t}$ is the fraction of nodes with degree k at time t
2. The probability that a new edge arrives to a node of degree k is
$$\frac{kp_{k,t}}{\sum_k kp_{k,t}} = \frac{kp_{k,t}}{2}$$
3. The change of the number of nodes of degree k in one time step
$$(t+1)p_{k,t+1} - tp_{k,t} = \begin{cases} \frac{1}{2}(k-1)p_{k-1,t} - \frac{1}{2}kp_{k,t} & \text{for } k > 1 \\ 1 - \frac{1}{2}p_{1,t} & \text{for } k = 1 \end{cases}$$
4. Stationary solution: $p_{k,t+1} = p_{k,t} := p_k$
5. The corresponding difference equation finally yields
$$p_{\text{eq}}(k) = \frac{4}{k(k+1)(k+2)} \sim \frac{1}{k^3}$$

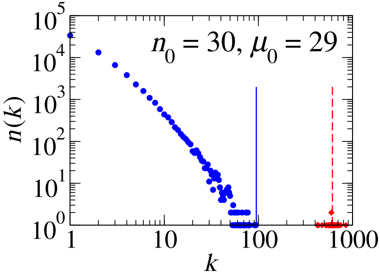
The basic PA model: advantage of the first

- Coming back to $\overline{k_i(t)} \approx \sqrt{t/i}$: the first nodes have by far the highest average degree
  - The power-law degree distribution solely due to the first nodes (no *"American dream"*)

Berset & Medo, EPJ B 86, 260, 2013

The basic PA model: advantage of the first

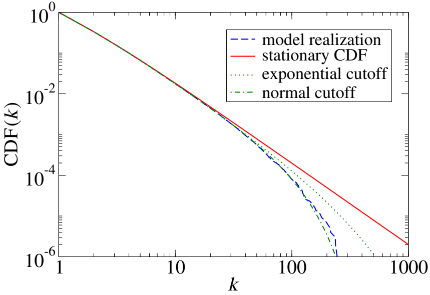
- Coming back to $\overline{k_i(t)} \approx \sqrt{t/i}$: the first nodes have by far the highest average degree
  - The power-law degree distribution solely due to the first nodes (no *"American dream"*)
- Worse still: infinite equlibration time for the degree distribution when there are several initial nodes



Berset & Medo, EPJ B 86, 260, 2013

The basic PA model: advantage of the first

- Coming back to $\overline{k_i(t)} \approx \sqrt{t/i}$: the first nodes have by far the highest average degree
  - The power-law degree distribution solely due to the first nodes (no *"American dream"*)
- Worse still: infinite equlibration time for the degree distribution when there are several initial nodes

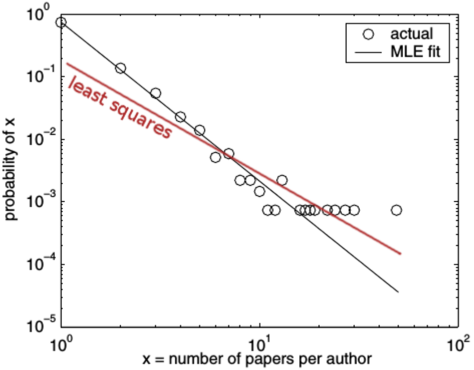


normal cutoff:
$P(k) \sim P_{\text{eq}}(k) \exp\left[-(k/\lambda)^2\right]$

Berset & Medo, EPJ B 86, 260, 2013

A detour: fitting (power-law) distribution

- Avoid fitting a straight line in the log-log plot



A detour: fitting (power-law) distribution

- Avoid fitting a straight line in the log-log plot
- A principled approach: Clauset et al, SIAM Review 51, 661, 2009
  - Key tools: maximum likelihood estimate, Kolmogorov-Smirnov statistic, *p*-values
- Advantages:
  - A better estimate of the exponent value:

$$\hat{\alpha} = 1 + n\left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}}\right]^{-1}$$ (exact when x_i's are continuous)

A detour: fitting (power-law) distribution

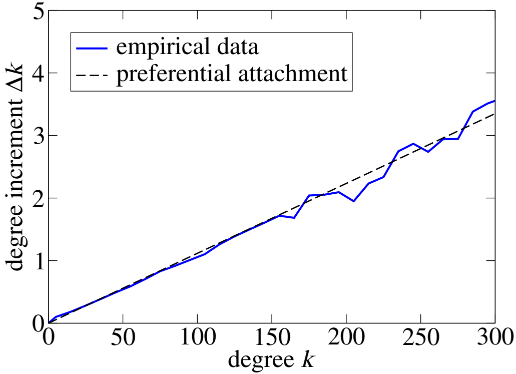
- Avoid fitting a straight line in the log-log plot
- A principled approach: Clauset et al, SIAM Review 51, 661, 2009
  - Key tools: maximum likelihood estimate, Kolmogorov-Smirnov statistic, *p*-values
- Advantages:
  - A better estimate of the exponent value:

$$\hat{\alpha} = 1 + n\left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}}\right]^{-1}$$ (exact when x_i's are continuous)

  - A way to estimate $x_{\min}$ and the cutoff parameter λ (if appropriate)
  - More generally: a way to compare between different fitting distributions
    - It's easy to mistake a log-normal distribution for power-law
    - In fact, a power law is rarely the best option

PA in scientific citation data

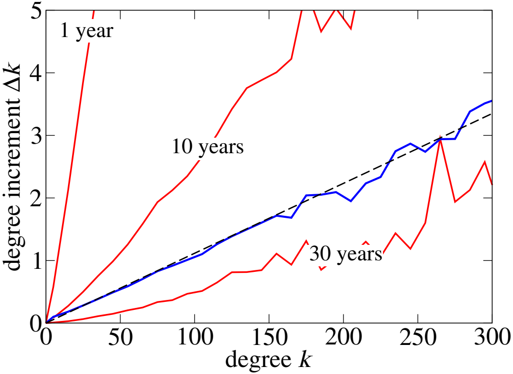
Journals of the American Physical Society from 1893 to 2009:



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

PICKED UP AFTER 8 HOURS

FRANKLIN HOPEFUL ALL DAY

HEAD OF THE LINE ABOARD

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

RESCUERS THERE TOO LATE

WOMEN AND CHILDREN FIRST

SEA SEARCH FOR OTHERS

OLYMPIC SENDS THE NEWS

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
  - When $D_R(t) \to 0$, the popularity of nodes eventually saturates

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
  - When $D_R(t) \to 0$, the popularity of nodes eventually saturates
- 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)

Solving the relevance model

$$P(i,t) = \frac{k_i(t)R_i(t)}{\sum_{j=1}^{t} k_j(t)R_j(t)} = \frac{k_i(t)R_i(t)}{\Omega(t)}$$

Solving the relevance model

$$\frac{\mathrm{d}\overline{k_i(t)}}{\mathrm{d}t} \approx P(i,t) = \frac{k_i(t)R_i(t)}{\sum_{j=1}^{t} k_j(t)R_j(t)} = \frac{k_i(t)R_i(t)}{\Omega(t)} \approx \Omega^*$$

Solving the relevance model

$$\frac{d\overline{k_i(t)}}{dt} \approx P(i,t) = \frac{k_i(t)R_i(t)}{\sum_{j=1}^{t} k_j(t)R_j(t)} = \frac{k_i(t)R_i(t)}{\Omega(t)} \approx \Omega^*$$

$$\Downarrow$$

$$\overline{k_i^F} = \exp\left(\frac{1}{\Omega^*}\int_0^\infty R_i(t)\,dt\right) = \exp\left(T_i/\Omega^*\right)$$

Solving the relevance model

$$\frac{\mathrm{d}\overline{k_i(t)}}{\mathrm{d}t} \approx P(i,t) = \frac{k_i(t)R_i(t)}{\sum_{j=1}^{t} k_j(t)R_j(t)} = \frac{k_i(t)R_i(t)}{\Omega(t)} \approx \Omega^*$$

$$\Downarrow$$

$$\overline{k_i^F} = \exp\left(\frac{1}{\Omega^*}\int_0^\infty R_i(t)\,\mathrm{d}t\right) = \exp\left(T_i/\Omega^*\right)$$

- $R_i(t)$ matters little, it's T_i what's important
- Ω^* can be set to achieve the desired $\langle k \rangle$
- Very strong (exponential) dependence between T and popularity

Fitness and aging: conclusions

- PA with fitness and aging as a relevant model for *information* networks
- There are many possible applications. . .
- 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

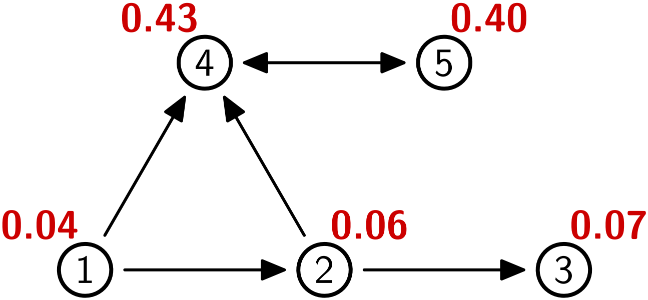
- PageRank is essentially a node centrality (importance) measure
- Simplest centrality: degree (counting the links—local)
- Non-local: PageRank (counting weighted links)
- Assign score $p_i^{(t)}$ to each node which initially is uniform: $p_i^{(0)} = 1/N$

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

  - $j \to i$ 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 ($c = 0$: 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

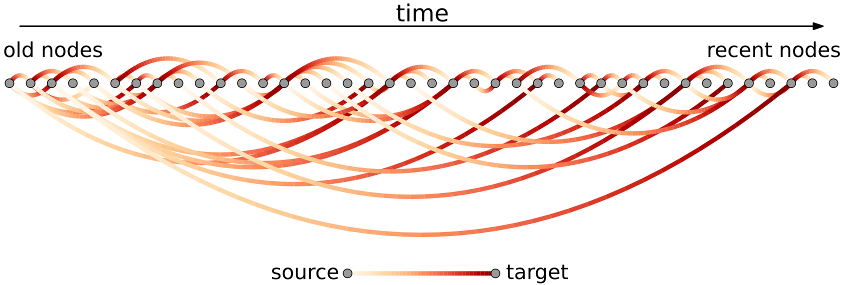
- The decay of relevance: $D_R(t)$
  - Node relevance influences the in-coming links

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

Two models 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 to create an outgoing link is

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

Two models 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 to create an outgoing link is

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

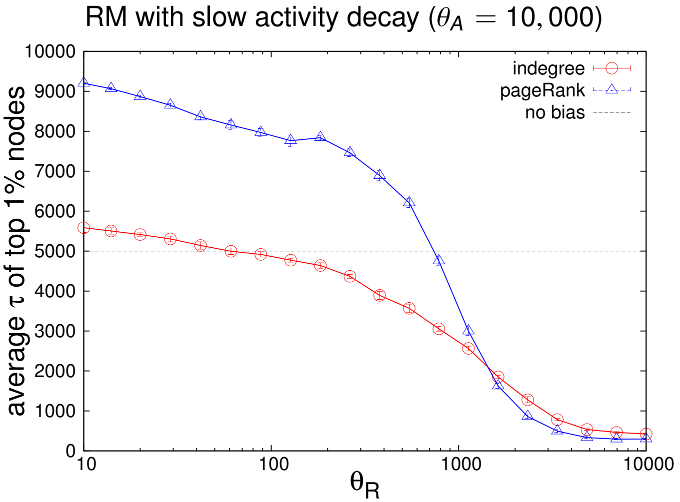
1. Relevance model (RM)

$$P_i^{in}(t) \sim (k_i^{in}(t) + 1)\, f_i\, D_R(t - \tau_i)$$

2. Extended fitness model (EFM)

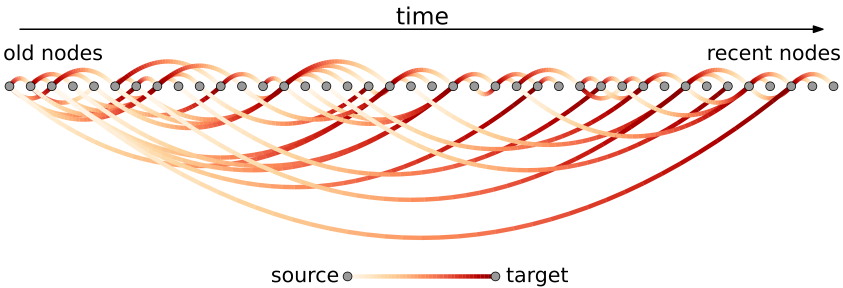
$$P_{i;j}^{in}(t) \sim (k_i^{in}(t) + 1)^{1-f_j}\, f_i^{f_j}\, D_R(t - \tau_i)$$

The biases of PageRank

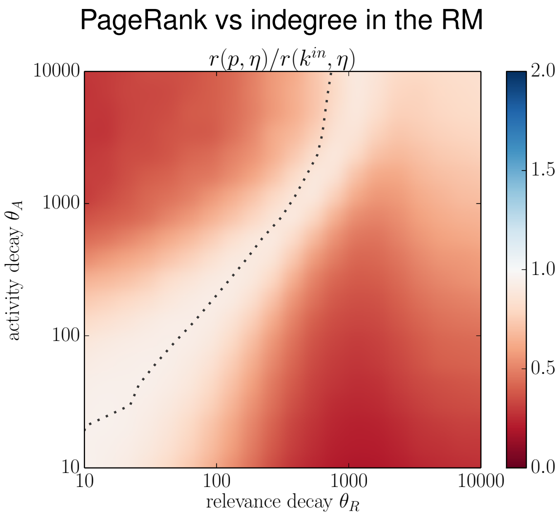


The biases of PageRank

Why the new kind of bias?

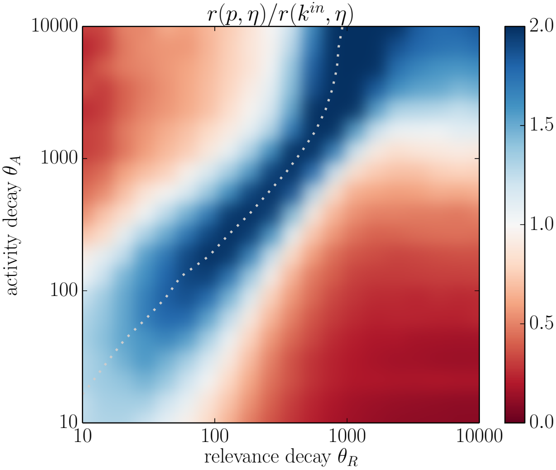


The biases of PageRank



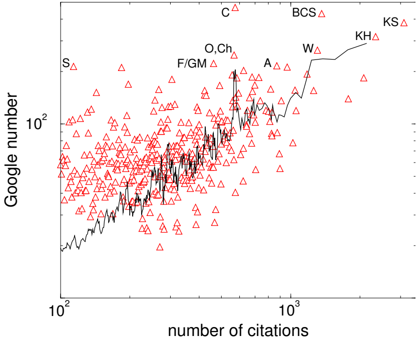
The biases of PageRank

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…

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

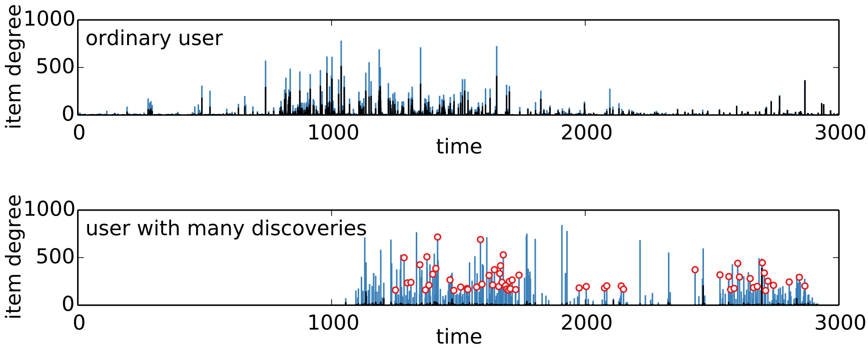
- Bipartite user-item data (e.g., *who* bought *what* at Amazon.com)
  - Similar behavior in monopartite social data (user-user)
- Previous research shows/assumes that users are driven by popularity combined with fitness and aging

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 informaton 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?

- Under the null hypothesis, we can generate the number of discoveries at will

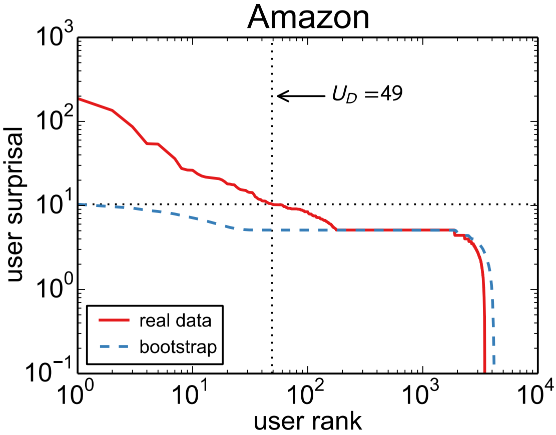
Algorithm 1 Using bootstrap to find the average highest user surprisal

Run many times
  Go over all users
    Draw d_i from the binomial distribution
    Compute the corresponding s_i
  Find the highest surprisal value
Report the average highest surprisal value

See C. R. Shalizi, The Bootstrap, American Scientist (2010) for more details on bootstrap

Discoverer or a lucky guy?

- Under the null hypothesis, we can generate the number of discoveries at will

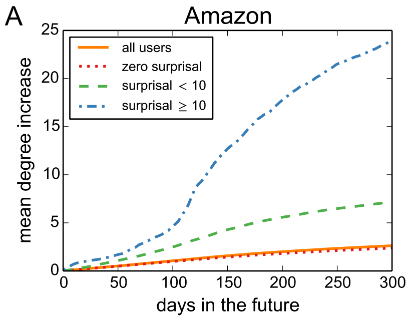


But. . . Is this any useful?

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

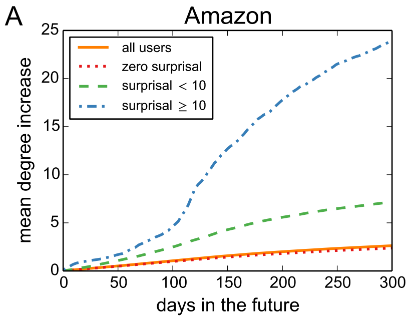
But. . . Is this any useful?

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



But. . . 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



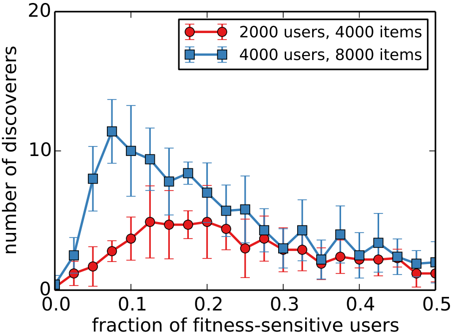
The answer: Yes, potentially very useful!

The game changer

- 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: $f_i(t)D_R(t)$

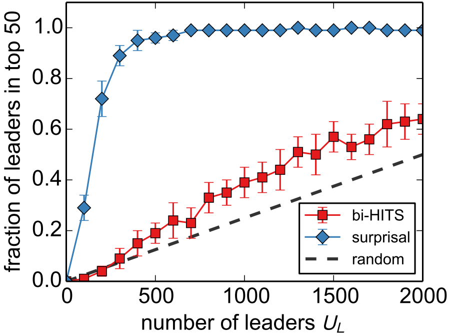
The game changer

- 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



The game changer

- 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 poses a puzzle to classical ranking algorithms



The game changer

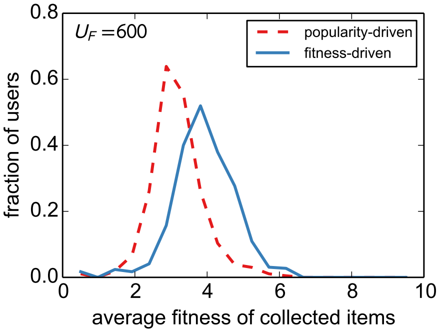
- Network growth model with to rules reproduces the real data patterns
  1. Some users are popularity-driven: $k_i(t)D_R(t)$
  2. Others are fitness-driven: $f_i(t)D_R(t)$
- Model data poses a puzzle to classical ranking algorithms

Oh, and this bi-HITS thing there. . .

- HITS: a close sibling of PageRank (9,000+ citations)
- Original HITS: each node has two kinds of score
- Bi-HITS for bipartite networks:
  - Good items are those that are collected by many good users
  - Good users are those who collect good items

The game changer

- 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 poses 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 monoparite 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, Volume 2" (Birkhäuser-Springer, 2013)
4. M. Medo, Statistical validation of high-dimensional models of growing networks, Physical Review E 89, 032801, 2014
5. M. Medo, M. S. Mariani, A. Zeng, Y.-C. Zhang, Identification and modeling of discoverers in online social systems, arXiv:1509.01477
6. M. S. Mariani, M. Medo, Y.-C. Zhang, Ranking nodes in growing networks: When PageRank fails, arXiv:1509.01476