Evolving networks: models and applications

Matúš Medo

University of Fribourg, Switzerland

Advanced theory of complex networks

Lectures 5 & 6 (April 7, 2016), IMT Lucca

Outline

1. Growing networks
   - Models of information networks with aging
   - Models of social networks (???)
2. Applications
   - Forecasting the popularity of research papers
   - Quantifying the significance of scientific papers
   - Finding the users who “know better” in e-commerce systems

Goals:
1) Understand how networks grow
2) Understand how this understanding can be useful

Part 1

Models of growing networks

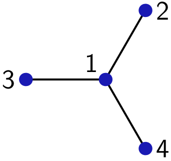
Preferential attachment (PA)

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

Preferential attachment (PA)

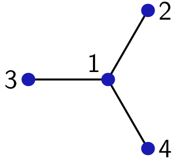
- A classical network model
  - Yule (1925), Simon (1955), Price (1976), Barabási & Albert (1999)
  - Growth of cities, citations of scientific papers, WWW,. . .
- Nodes and links are added with time
- Probability that a node acquires a new link proportional to its current degree

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



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

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

- The master equation can do much more…

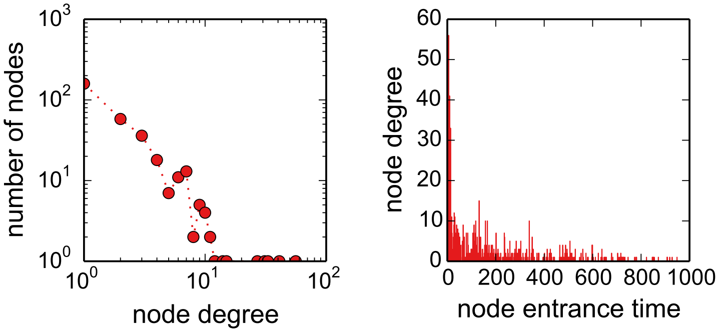
The basic PA model: advantage of the first

- 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”*)

The basic PA model: advantage of the first

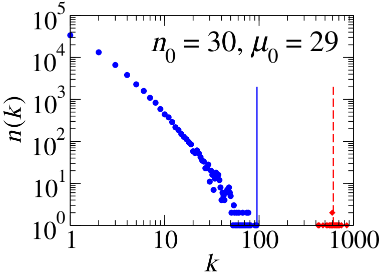
- Coming back to $\overline{k_i(t)} \approx \sqrt{t/i}$: the first nodes have by far the highest average degree
  - The power-law degree distribution solely due to the first nodes (no *“American dream”*)

sample network for $N = 1000$



The basic PA model: advantage of the first

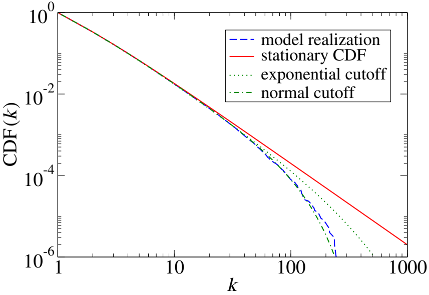
- Coming back to $\overline{k_i(t)} \approx \sqrt{t/i}$: the first nodes have by far the highest average degree
  - The power-law degree distribution solely due to the first nodes (no *“American dream”*)
- Even worse: infinite equlibration time for the degree distribution when there are several initial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

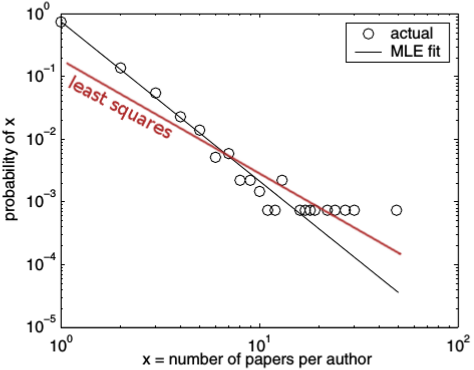


normal cutoff:
$$P(k) \sim P_{\mathrm{eq}}(k)\,\exp\left[-(k/\lambda)^2\right]$$

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:

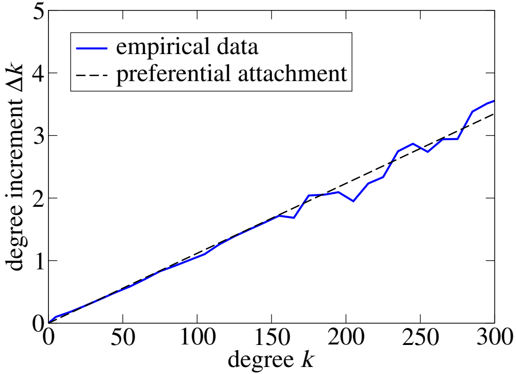
$$\hat{\alpha} = 1 + n\left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}}\right]^{-1}$$

(exact when x_i's are continuous)

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

PA in scientific citation data

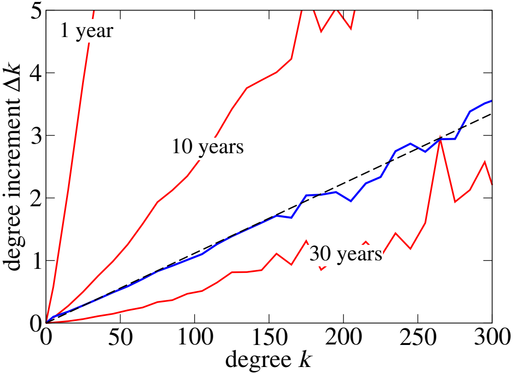
Journals of the American Physical Society from 1893 to 2009:



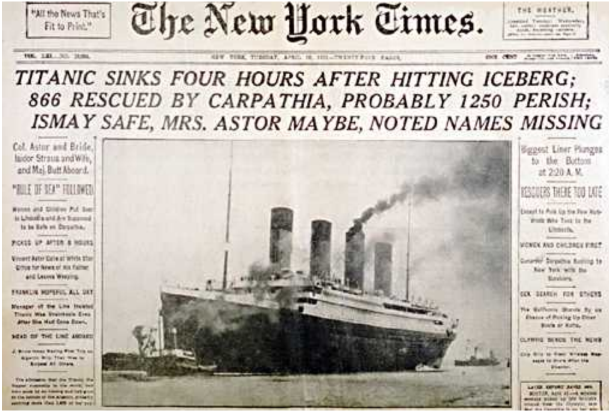
See also Adamic & Huberman (2000), Redner (2005), Newman (2009),. . .

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. [illegible]

VOL. LXI...NO. [illegible] NEW YORK, TUESDAY, APRIL 16, 1912—TWENTY-FOUR PAGES. ONE CENT [illegible]

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

Women and Children Put Over in Lifeboats and Are Supposed to be Safe on Carpathia.

PICKED UP AFTER 8 HOURS

Vincent Astor Calls at White Star Office for News of His Father and Leaves Weeping.

FRANKLIN HOPEFUL ALL DAY

Manager of the Line Insisted Titanic Was Unsinkable Even After She Had Gone Down.

HEAD OF THE LINE ABOARD

J. Bruce Ismay Making First Trip on Gigantic Ship That Was to Surpass All Others.

[illegible]

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

RESCUERS THERE TOO LATE

Except to Pick Up the Few Hundreds Who Took to the Lifeboats.

WOMEN AND CHILDREN FIRST

Cunarder Carpathia Rushing to New York with the Survivors.

SEA SEARCH FOR OTHERS

The California Stands By on Chance of Picking Up Other Boats or Rafts.

OLYMPIC SENDS THE NEWS

Only Ship to Flash Wireless Messages to Shore After the Disaster.

LATER REPORT SAVES 868.

BOSTON, [illegible]

Growing information networks: they are everywhere

- Citations among scientific papers
  - Directed monopartite network
  - Outgoing links (references) are fixed
- The World Wide Web
  - Directed monopartite network
  - New nodes and new links are added gradually
- E-commerce data
  - Undirected bipartite network
  - Links connect users with the items that they have purchased/rated/collected
- Wikipedia
  - Semantic centrality of page topic important

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

Comparing network growth models (Phys Rev E 89, 032801, 2014)

- Between models, often only aggregate results are compared
- Detailed look: likelihood that data has been produced by the model

Likelihood (probability)

Go over all edges

$$\mathcal{L}(\mathcal{D}|\mathcal{M}) = \prod_{i=1}^{E} P(n_i, t_i|\mathcal{M})$$

Data

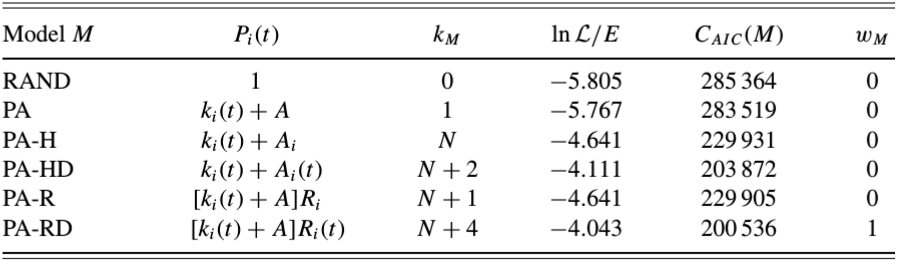
Model

Link to node n_i

At time t_i

Comparing network growth models (Phys Rev E 89, 032801, 2014)

- Between models, often only aggregate results are compared
- Detailed look: likelihood that data has been produced by the model



Model M	$P_i(t)$	k_M	$\ln \mathcal{L}/E$	$C_{AIC}(M)$	w_M
RAND	1	0	-5.805	285 364	0
PA	$k_i(t) + A$	1	-5.767	283 519	0
PA-H	$k_i(t) + A_i$	N	-4.641	229 931	0
PA-HD	$k_i(t) + A_i(t)$	$N + 2$	-4.111	203 872	0
PA-R	$[k_i(t) + A]R_i$	$N + 1$	-4.641	229 905	0
PA-RD	$[k_i(t) + A]R_i(t)$	$N + 4$	-4.043	200 536	1

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

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

$$\Downarrow$$

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

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

Growth of social networks: typical features

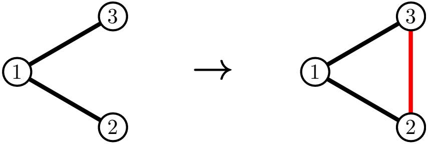
1. High clustering (a friend of a friend is a friend)
2. In directed networks, high reciprocity (Flickr: about 70%)
3. Broad degree distribution (no surprise)
4. Small average distance (six degrees of separation)
5. Assortativity (“Birds of a feather flock together”)

How to get: high clustering

- Due to constraints:
  - If for example, nodes are distributed in real space and the probability of being connected decreases with distance
  - If $A \sim B$ and $B \sim C$, probably A is close to B and B is close to C
  - Hence A is also close to C and they are likely to be connected
  - See Boguñá et al, Phys. Rev. E 70, 056122 (2004)

How to get: high clustering

- Due to constraints:
  - If for example, nodes are distributed in real space and the probability of being connected decreases with distance
  - If $A \sim B$ and $B \sim C$, probably A is close to B and B is close to C
  - Hence A is also close to C and they are likely to be connected
  - See Boguñá et al, Phys. Rev. E 70, 056122 (2004)
- Triangle closing:
  - Imagine that links are introduced with time (growth)
  - The growth rule is conditional—with certain probability a random link is added, otherwise a "triangle" is chosen at random and "closed"



How to get: . . .

- Link reciprocity
  - Easy: every user has some probability to reciprocate the incoming links
- Broad degree distribution
  - Preferential attachment is again an option (for both out- and in-degree)
  - Inherent node feature (node activity for out-degree)
- Small average distance
  - Due to a mixture of short-range order (lattice) and long-range randomness (shortcuts)
  - The seminal model by Watts & Strogatz (Nature 393(6684), 440, 1998)
- Assortativity
  - As opposed to technological networks, social networks are mostly assortative (with respect to degree or other node properties)
  - In growth, we can prefer linking to nodes of similar degree: $P(k_2|k_1) \sim e^{-|k_2-k_1|}$, for example

Example model: Kumar, Novak, Tomkins (2006)

- Three kinds of users: passive, inviter, linker
  1. Passive users: targets for the others
  2. Inviters: bring new users to the networks
  3. Linkers: users who actively seek new contacts in the network

Example model: Kumar, Novak, Tomkins (2006)

- Three kinds of users: passive, inviter, linker
  1. Passive users: targets for the others
  2. Inviters: bring new users to the networks
  3. Linkers: users who actively seek new contacts in the network

- Nodes and links are gradually added, the proportion P:I:L is given
- In each time step, a new node and ε links are added
- User *i* becomes active with probability $k_i + 1$
- Active Inviter links to an additional new Passive node
- Active Linker prefers linking to other linkers

Example model: Kumar, Novak, Tomkins (2006)

- Three kinds of users: passive, inviter, linker
  1. Passive users: targets for the others
  2. Inviters: bring new users to the networks
  3. Linkers: users who actively seek new contacts in the network

- Nodes and links are gradually added, the proportion P:I:L is given
- In each time step, a new node and ε links are added
- User *i* becomes active with probability $k_i + 1$
- Active Inviter links to an additional new Passive node
- Active Linker prefers linking to other linkers

- Flickr: $p = (0.25, 0.35, 0.40)$, $\varepsilon = 6$, preference for linkers $\gamma = 15$
  - This reproduces well the Flickr social network

Network growth models: summary

- Key ingredients for information networks:
  1. Preferential attachment
  2. (Heterogeneous) Node fitness
  3. Aging

- To reproduce well social networks, we need many bits. . .

Network growth models: summary

- Key ingredients for information networks:
  1. Preferential attachment
  2. (Heterogeneous) Node fitness
  3. Aging

- To reproduce well social networks, we need many bits. . .

- Models establish a playground!

Application 1

Quantifying Long-Term Scientific Impact

Forecasting the citation count

- Citations are commonly used to measure a paper’s importance
  - They also serve as input for other metrics (*h*-index for researchers, impact factor for journals, etc.)
- But: citations take long time to accumulate
- Can we predict the eventual citation count of a paper?
- Wang et al (Science 342, 6154, pp. 127, 2013) say yes!

How to forecast

- Observation: Relevance of papers has a log-normal form

$$R_i(t) = \eta_i \times \frac{1}{\sqrt{2\pi}\sigma_i t} \exp\left[- \frac{(\ln t - \mu_i)^2}{2\sigma_i^2} \right]$$

  - Here η_i is fitness of paper i, μ_i is its immediacy, and σ_i is its longevity

How to forecast

- Observation: Relevance of papers has a log-normal form

$$R_i(t) = \eta_i \times \frac{1}{\sqrt{2\pi}\sigma_i t} \exp\left[- \frac{(\ln t - \mu_i)^2}{2\sigma_i^2} \right]$$

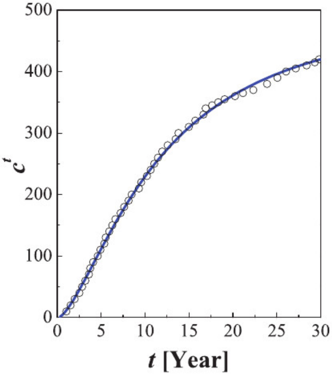
  - Here η_i is fitness of paper i, μ_i is its immediacy, and σ_i is its longevity

- Integrating the master equation, we get

$$k_i(t) = m\left[e^{\lambda_i \Phi\left(\frac{\ln t - \mu_i}{\sigma_i}\right)} - 1 \right] \qquad (*)$$

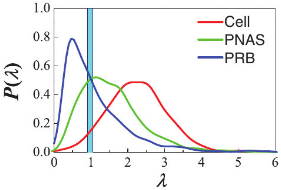
  where m is the number of references per paper, β is the growth rate of the number of papers per time unit, A is a normalization constant, $\lambda_i = \eta_i\beta/A$ is paper's relative fitness, and $\Phi(\cdot)$ is the CDF of the standard normal distribution

Forecasting results

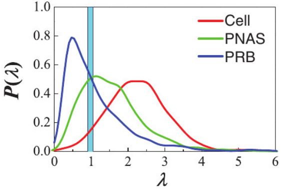


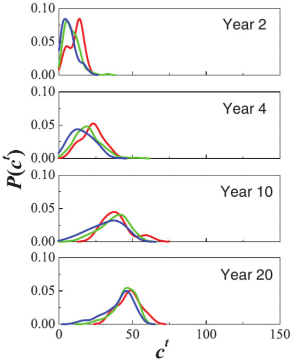
Cumulative citation count in the APS data and their fit with ($*$)

Forecasting results

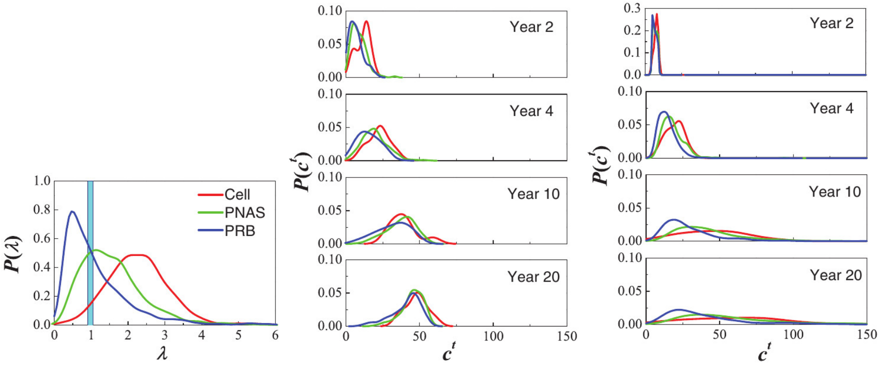


Forecasting results



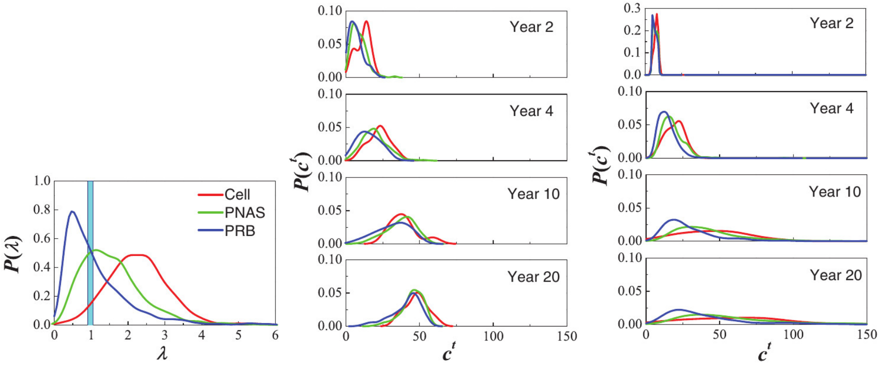


Forecasting results



Bottom line: in the long term, λ predicts citations better than the short-term citation count

Forecasting results



But: careful fitting needed (regularization?), $\gtrsim$ 10 years to predict individual papers well

Application 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

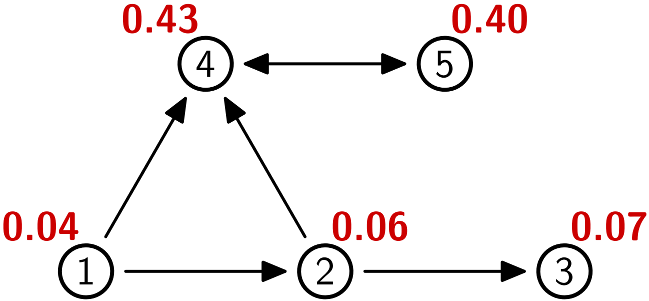
- PageRank is essentially a node centrality (importance) measure
- Simplest centrality: degree (counting the links—local)
- Non-local: PageRank (counting weighted links)
- Assign score $p_i^{(t)}$ to each node which initially is uniform: $p_i^{(0)} = 1/N$

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

  - $j \to i$ 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

- Decay of relevance: $D_R(t) = \exp(-t/\theta_R)$
  - Node relevance influences the in-coming links

Two forms of aging in information networks

- Decay of relevance: $D_R(t) = \exp(-t/\theta_R)$
  - Node relevance influences the in-coming links
- Decay of activity: $D_A(t) = \exp(-t/\theta_A)$
  - Nodes activity influences the out-going links

Two forms of aging in information networks

- Decay of relevance: $D_R(t) = \exp(-t/\theta_R)$
  - Node relevance influences the in-coming links
- Decay of activity: $D_A(t) = \exp(-t/\theta_A)$
  - Nodes activity influences the out-going links

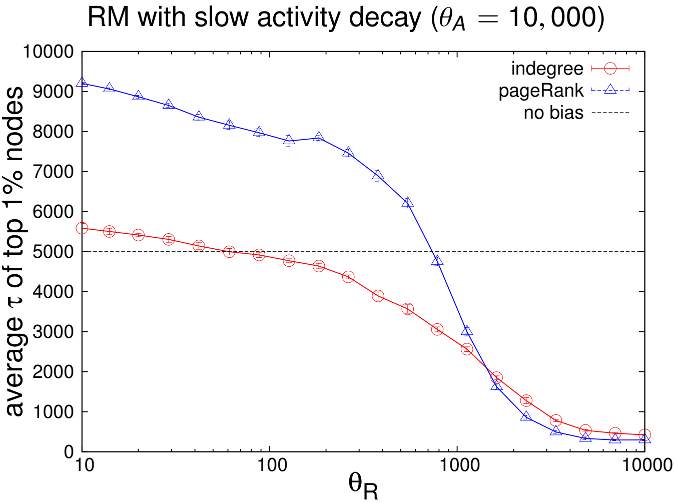
- The probability to attract an incoming link

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

- The probability to create an outgoing link

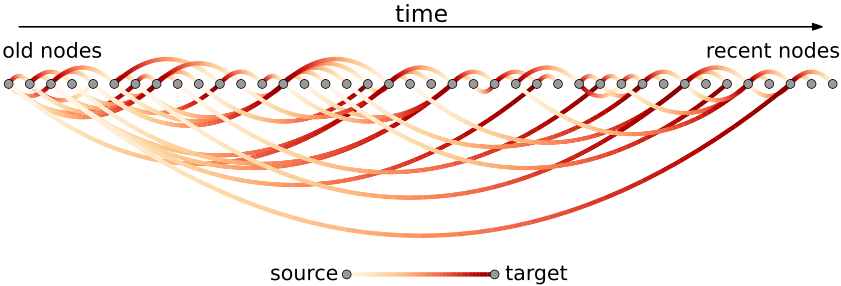
$$P_i^{out} \sim A_i D_A(t - \tau_i)$$

The biases of PageRank



The biases of PageRank

Why the new kind of bias?



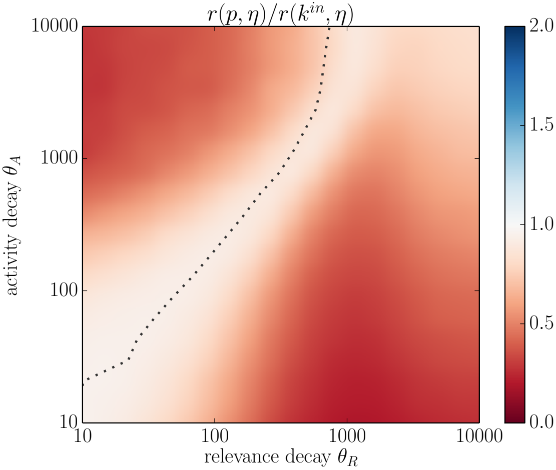
The biases of PageRank

Now the question is:
Does PageRank rank the nodes well?

- Recall, quality of node i is represented by its fitness f_i
- Hence, compare $r(p_i, f_i)$ with $r(k_i, f_i)$ to see whether PageRank outperforms in-degree

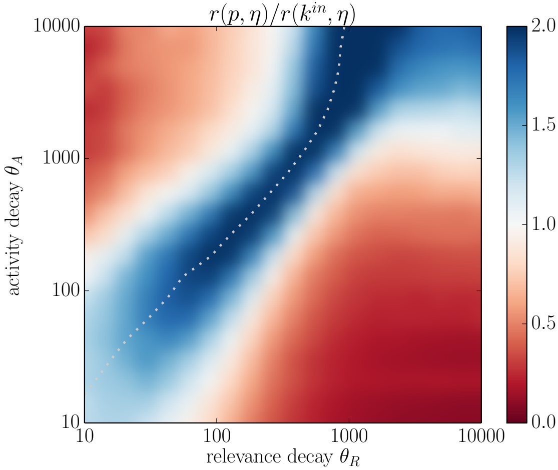
The biases of PageRank

PageRank vs indegree in the RM



The biases of PageRank

PageRank vs indegree in the EFM



Application 3

Correcting the bias of PageRank

Why bias is bad

> Metrics: usually well intentioned, not always well informed, often ill applied
>
> Leiden Manifesto (2015)

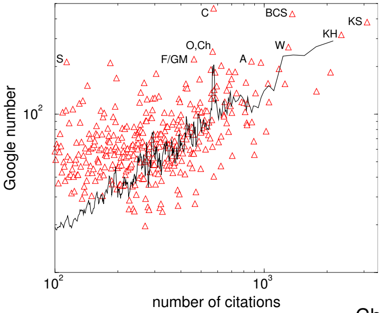
Why bias is bad

> Metrics: usually well intentioned, not always well informed, often ill applied
>
> Leiden Manifesto (2015)

- In citation data, Θ_R is a few years and $\Theta_A = 0$
- We expect PageRank to perform very badly, yet it is often applied here

Why bias is bad

> Metrics: usually well intentioned, not always well informed, often ill applied
>
> Leiden Manifesto (2015)



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

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

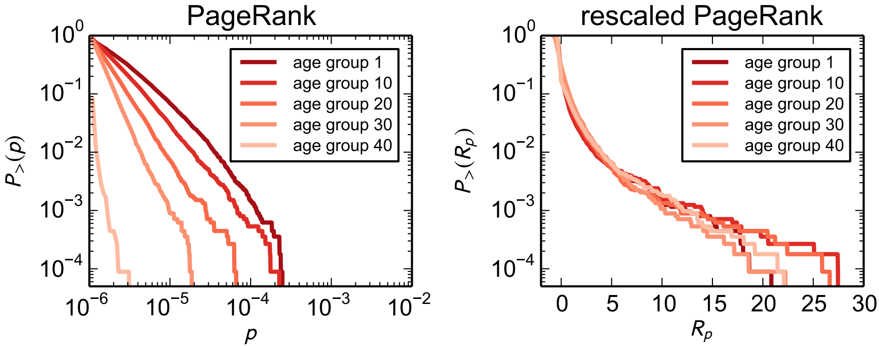
Correcting PageRank

- 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`)

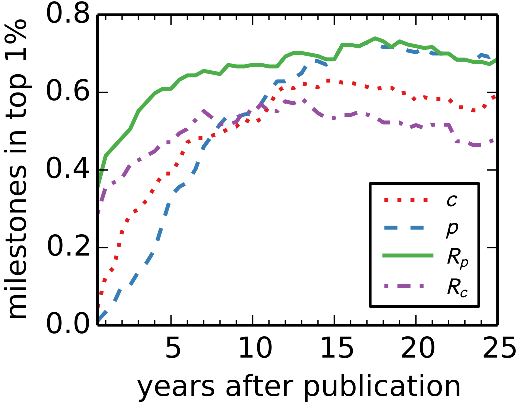
Rescaled PageRank: results



↙

Allows us to fairly compare all papers!

Rescaled PageRank: results



Note: CiteRank is competitive with R_p in some aspects

Application 4

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

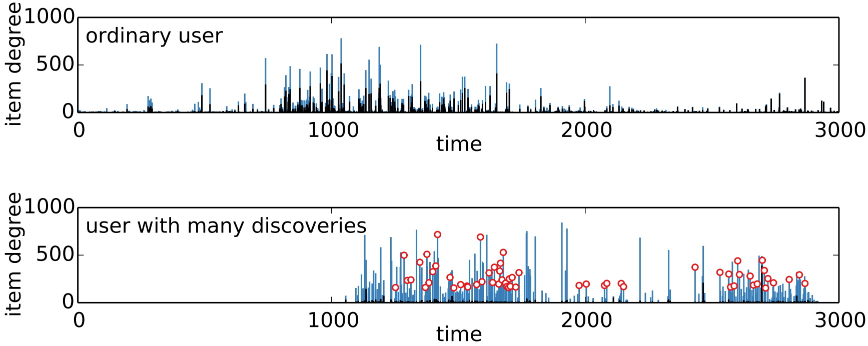
But: is this the whole story?

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

- We know the number of links k_i and the number of discoveries d_i for each user
- Which users are truly exceptional?

How to quantify the user success

- We know the number of links k_i and the number of discoveries d_i for each user
- Which users are truly exceptional?
- The overall discovery probability is $p_D = D/L = (\sum_i d_i)/(\sum_i k_i)$
- Assuming that all users and links are equal, the probability that a user makes at least d_i discoveries in k_i attempts is

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

Top users in the Amazon data

Rank	k_i	d_i	r_i	P_i^0	$s_i = -\ln P^0$
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	$s_i = -\ln P^0$
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
			...		

Is this not just luck?

Discoverer or a lucky guy?

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

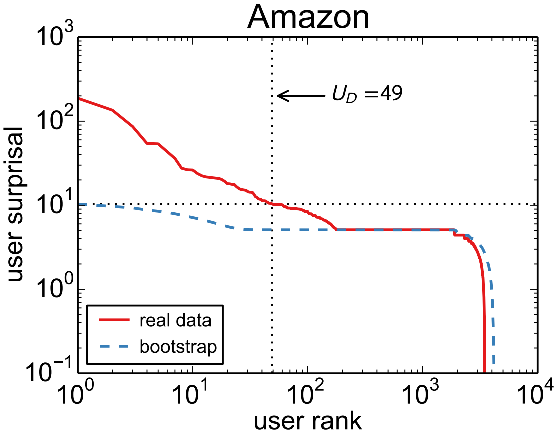
Algorithm 1 Using bootstrap to find the average highest user surprisal

```
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

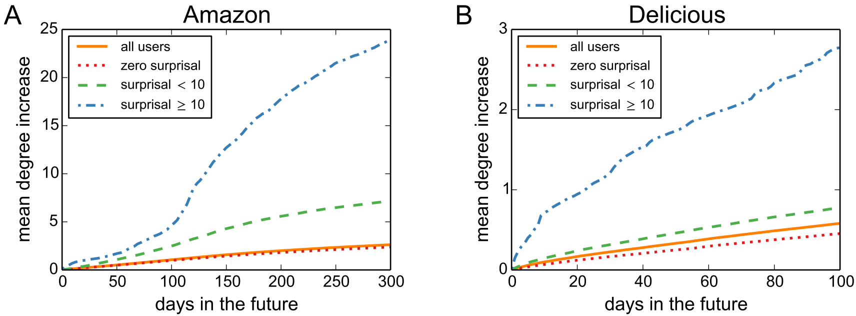


But. . . Is this any useful?

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

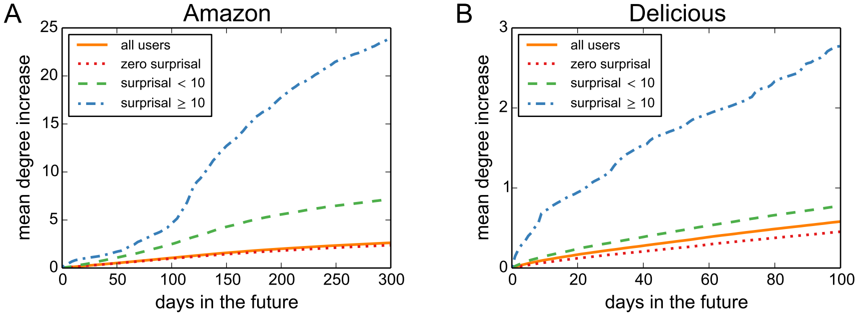
But. . . Is this any useful?

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!

Discoverers: conclusions

- We find discoverers in almost any information network we look at
- The surprisal measure is not biased by activity and efficiently identifies the unusual users
- Once identified, discoverers can be used for prediction
- 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?

General lessons

- We know a lot about the evolution of complex systems
- **Let the data drive you**
- Do not use (trust in) wrong models
- Do not use wrong algorithms
- Do not use right algorithms wrongly

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