

Complexity insights into information filtering

Matúš Medo

University of Fribourg, Switzerland

ICT Applications to Non-Equilibrium Social Sciences

11-12 June, 2013, Lisbon

Outline

- 1 Growth of information networks
- 2 Physics-motivated approach to recommendation
- 3 Crowd-avoidance in recommendation

Part 1

Growth of information networks



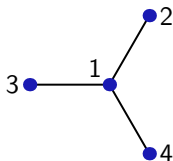
Preferential attachment

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

Preferential attachment

- 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

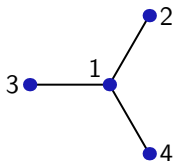
$$P(i, t) \sim k_i(t)$$



Preferential attachment

- 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

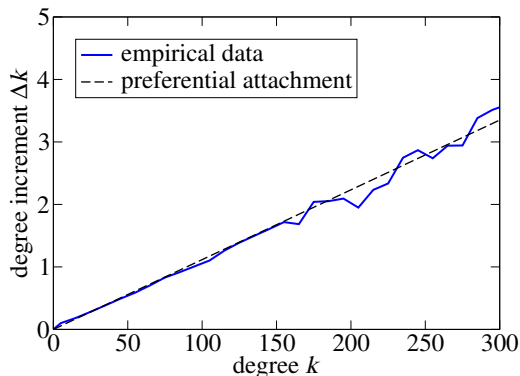
$$P(i, t) \sim k_i(t)$$



- Pros: simple, produces a power-law degree distribution

PA in scientific citation data

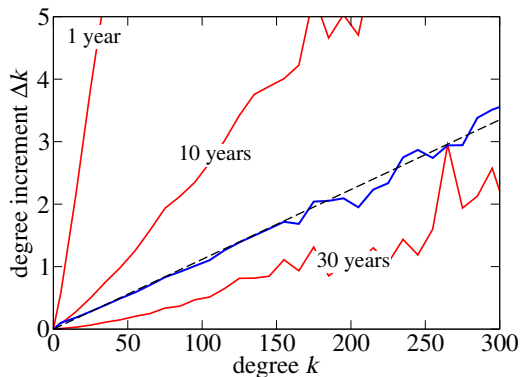
Journals of the American Physical Society from 1893 to 2009:



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 BEAST.

THE. 5.21. 1912. 30. 30. 30. NEW YORK, FRIDAY, APRIL 19, 1912. TWENTY-FIVE CENTS.

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, Isaac Strauss and Wife, and Maj. Butt Aboard.

"HOLE OF DEEP" FOLLOWED

Women and Children Put Save in Lifeboats and Are Supposed to Be Safe on Carpathia.

PICKED UP AFTER 8 HOURS

Survivor Taken to White Star Office for News of His Father and Lesser Wrecking.

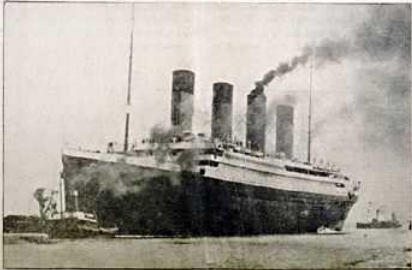
FRANKLIN HOSPITAL ALL SET

Manager of the Line Thinks There May Be Some Lives After Ship Had Gone Down.

HEAD OF THE LINE ANKERS

A Boat Seen Being Hoisted on Board of Ship.

The Atlantic has the Titanic, the biggest ship in the world, and the most of an island and the world, as the result of the American's... (text is small and partially obscured)



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

RESCUES THREE TOO LATE

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

WOMEN AND CHILDREN FIRST

General Serpahee Said to Have Taken with the Babies.

SEEK SEARCH FOR OTHERS

The Carpathia Starts by an Effort of Picking Up Other Boats or Rafts.

CLIPPING SENDS THE NEWS

Ship Will Be Hunted for Hours.

LARGE REPORT SAID 866

RESCUED. 1250 MAY BE PERISHING. MRS. ASTOR MAYBE. ISMAY SAFE. (text is small and partially obscured)

The model (PRL 107, 238701, 2011)

- Probability that node i attracts a new link

$$P(i, t) \sim \underbrace{k_i(t)}_{\text{degree}} \times \underbrace{R_i(t)}_{\text{relevance}}$$

- Relevance of every node decays with time
- When $R_i(t) \rightarrow 0$, the popularity of nodes eventually saturates

The model (PRL 107, 238701, 2011)

- Probability that node i attracts a new link

$$P(i, t) \sim \underbrace{k_i(t)}_{\text{degree}} \times \underbrace{R_i(t)}_{\text{relevance}}$$

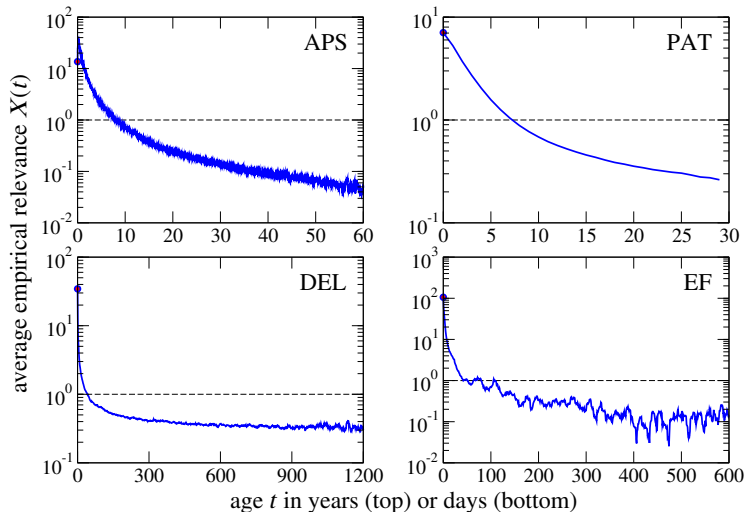
- Relevance of every node decays with time
 - When $R_i(t) \rightarrow 0$, the popularity of nodes eventually saturates
- **Good news:**
 - Produces various realistic degree distributions (power-law, etc.)

Datasets

- 1 Citations among papers published by the APS
- 2 Citations among the US patents
- 3 User collections of web bookmarks
- 4 Paper downloads from the Econophysics Forum

data description	label	nodes	links	span/resolution	Δt
APS citations	APS	450k	4.7M	117 years/daily	91 days
U.S. patents	PAT	3.2M	24M	31 years/yearly	1 year
web bookmarks	WEB	2.3M	4.2M	4 years/daily	10 days
paper downloads	EF	600	16k	23 months/daily	10 days

Decay of relevance



Future challenges

- Other properties of the model networks: clustering, community structure, . . .
 - Improved statistical validation
 - Troubles with high-dimensional statistics
 - Econophysics Forum data: new model is much more likely (10^{410} -times) than the second-best model (Eom-Fortunato, 2011)
-

- Knowledge of the dynamics can help select nodes that are most relevant now \implies useful tools?
- Theory says: Total relevance matters \implies useful tools?

Part 2

Physics-motivated approach to recommendation



Recommendation challenge

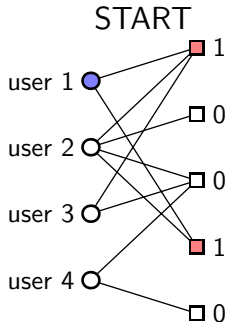
- There is often too much information available
 - Too many books to read, movies to watch, ...
 - How to choose?

Recommendation challenge

- There is often too much information available
 - Too many books to read, movies to watch, ...
 - How to choose?
- *Recommender systems* analyze data on past user preferences to predict possible future likes and interests
- Many approaches exist:
 - Collaborative filtering
 - Content-based analysis
 - Latent semantic models
 - Spectral methods
 - ...

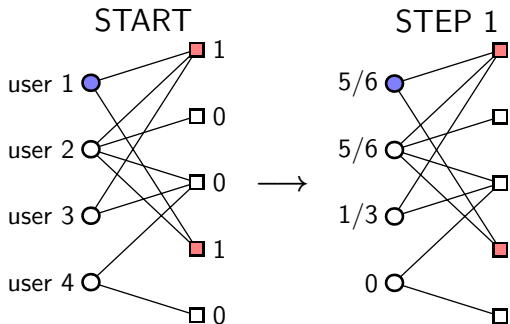
Recommendation by random walk

Two-step random walk:



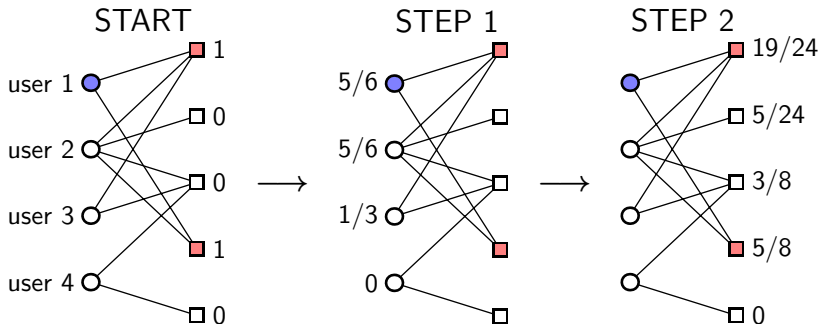
Recommendation by random walk

Two-step random walk:



Recommendation by random walk

Two-step random walk:



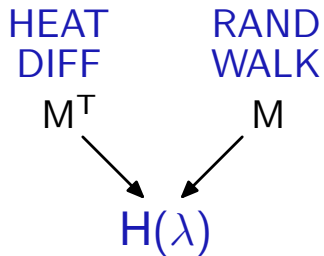
Key insights

- Random walk favors high-degree nodes
 - They have more ways to receive resources
- By contrast, heat diffusion favors low-degree nodes
 - If you touch many places, your temperature is likely to be average
 - If you touch a few places, you may get “lucky” and end hot

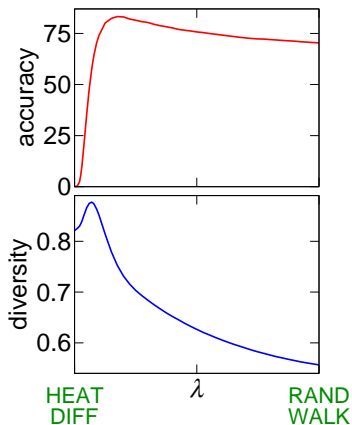
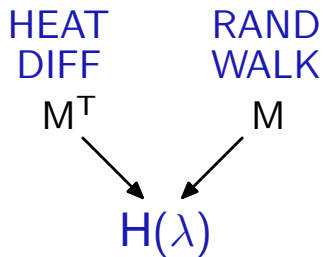
Key insights

- Random walk favors high-degree nodes
 - They have more ways to receive resources
- By contrast, heat diffusion favors low-degree nodes
 - If you touch many places, your temperature is likely to be average
 - If you touch a few places, you may get “lucky” and end hot
- Interestingly, the two processes are mathematically closely related
 - Their matrices are transpose of each other: M and M^T

Hybridization (PNAS 107, 4511, 2010)



Hybridization (PNAS 107, 4511, 2010)



Result: simultaneous improvement of accuracy **and** diversity

Part 3

Crowd-avoidance in recommendation

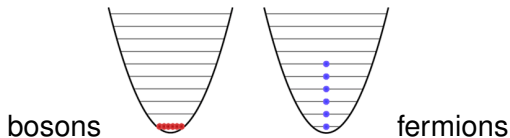


The challenge

- Traditional recommender systems do not care to how many users an item gets recommended
- Since they often have bias toward popularity, a small number of winners often emerges

The challenge

- Traditional recommender systems do not care to how many users an item gets recommended
- Since they often have bias toward popularity, a small number of winners often emerges
- This may be harmful:
 - Bar recommended to many people becomes overcrowded
 - In the long term, our information horizons shrink



Crowd-avoidance in recommendation (EPL 101, 20008, 2013)

- Easy to achieve ex post:

- 1 A recommendation algorithm produces a ranked list of items for each user
- 2 We impose maximal occupancy m for every item

An example with $m = 1$

user 1

item 7 ✓

item 3

Crowd-avoidance in recommendation (EPL 101, 20008, 2013)

■ Easy to achieve ex post:

- 1 A recommendation algorithm produces a ranked list of items for each user
- 2 We impose maximal occupancy m for every item

An example with $m = 1$

user 1	user 2
item 7 ✓	item 3 ✓
item 3	item 4

Crowd-avoidance in recommendation (EPL 101, 20008, 2013)

■ Easy to achieve ex post:

- 1 A recommendation algorithm produces a ranked list of items for each user
- 2 We impose maximal occupancy m for every item

An example with $m = 1$

user 1	user 2	user 3
item 7 ✓	item 3 ✓	item 7
item 3	item 4	item 2 ✓

Crowd-avoidance in recommendation (EPL 101, 20008, 2013)

- Easy to achieve ex post:

- 1 A recommendation algorithm produces a ranked list of items for each user
- 2 We impose maximal occupancy m for every item

An example with $m = 1$

user 1	user 2	user 3
item 7 ✓	item 3 ✓	item 7
item 3	item 4	item 2 ✓

- Two ways to enforce occupation:

- User-by-user (local optimization)
- Minimize the total rank of chosen items (global optimization)

Test case

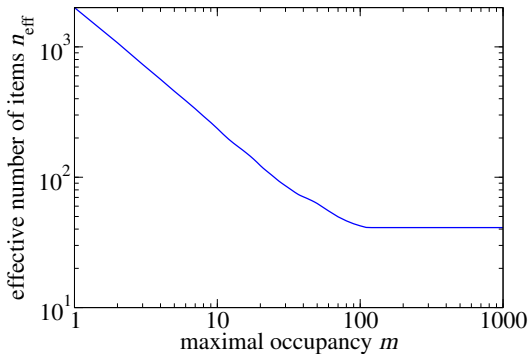
- Netflix subset with 2000 users
- One object recommended to each user ($m = 1$)

Test case

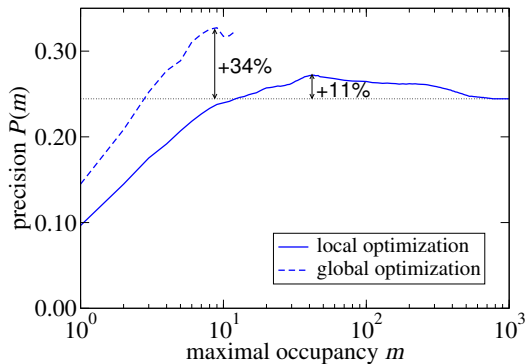
- Netflix subset with 2000 users
- One object recommended to each user ($m = 1$)
- There is no reason for real crowd avoidance in DVD rentals
- Country-production data would be a better test candidate but. . .

Crowd-avoidance enhances diversity

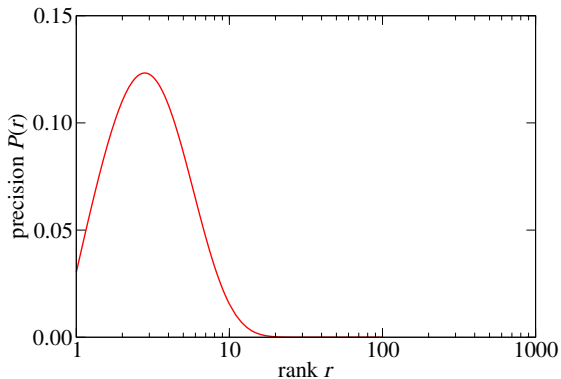
$$n_{\text{eff}} = \left(\sum_{\alpha} k_{\alpha} \right)^2 / \sum_{\alpha} k_{\alpha}^2$$



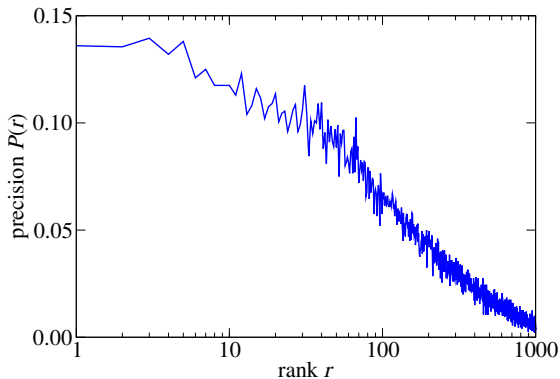
Crowd-avoidance enhances accuracy



Is it because our recommendations are wrong?



Is it because our recommendations are wrong?



No

The (probably) correct explanation

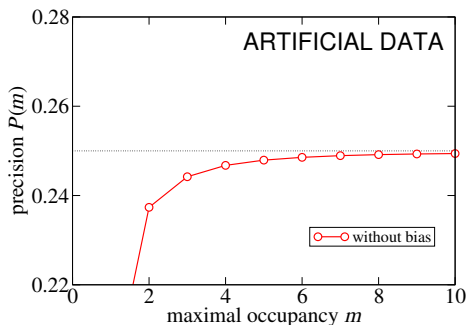
- Recommendation algorithms are often biased
 - Typically towards popular items but less tangible reasons exist too

The (probably) correct explanation

- Recommendation algorithms are often biased
 - Typically towards popular items but less tangible reasons exist too
- Hypothesis: crowd-avoidance suppresses these biases

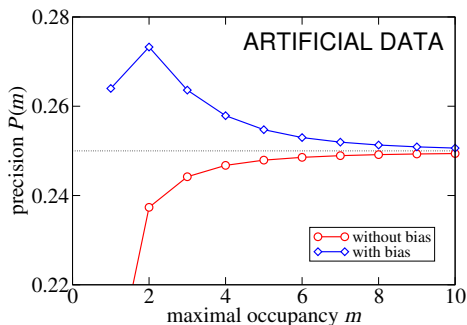
The (probably) correct explanation

- Recommendation algorithms are often biased
 - Typically towards popular items but less tangible reasons exist too
- Hypothesis: crowd-avoidance suppresses these biases



The (probably) correct explanation

- Recommendation algorithms are often biased
 - Typically towards popular items but less tangible reasons exist too
- Hypothesis: crowd-avoidance suppresses these biases



Crowd-avoidance: Summary

- Crowd-avoidance can improve both accuracy and diversity of recommendation
- A rare case where constraints improve the outcome

Crowd-avoidance: Summary

- Crowd-avoidance can improve both accuracy and diversity of recommendation
- A rare case where constraints improve the outcome

To do

- How best to introduce occupancy constraints?
 - Constraints heterogeneous over objects
 - Approaches to global optimization
- Study real data with crowd avoidance
 - Country-production data

Thank you for your attention!

Questions?