Identification, modeling and impact of discoverers in e-commerce systems

Matúš Medo

University of Fribourg, Switzerland

2016 Conference on Complex Systems, Amsterdam

19-22 September 2016, Torino

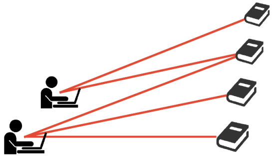
E-commerce systems and their users

- E-commerce systems: Amazon, Netflix, YouTube, . . .
- Users can:
  1. Buy things
  2. Consume content
     (*e.g.*, watch videos)
  3. Contribute content
  4. . . .

E-commerce systems and their users

- E-commerce systems: Amazon, Netflix, YouTube, ...
- Users can:
  1. Buy things
  2. Consume content (*e.g.*, watch videos)
  3. Contribute content
  4. ...

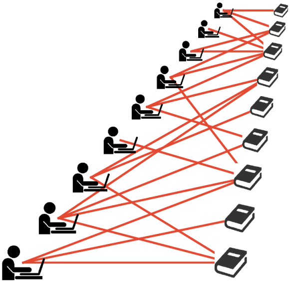
users grow a bipartite user-item network



E-commerce systems and their users

- E-commerce systems: Amazon, Netflix, YouTube, ...
- Users can:
  1. Buy things
  2. Consume content (*e.g.*, watch videos)
  3. Contribute content
  4. ...

users grow a bipartite user-item network



How to model this

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

How to model this

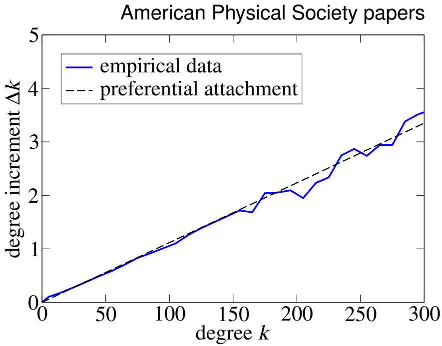
- Preferential attachment
  - Yule (1925), Simon (1955), Price (1976), Barabási & Albert (1999)
- Probability that item i attracts a new link:

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

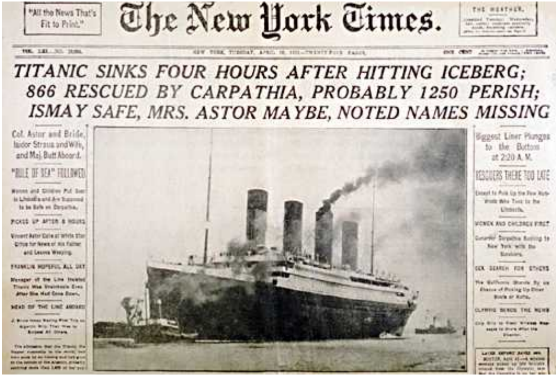
How to model this

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

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



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

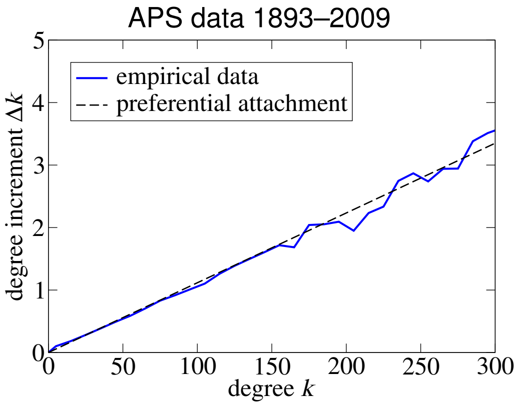
Biggest Liner Plunges to the Bottom at 2:20 A. M.

RESCUERS THERE TOO LATE

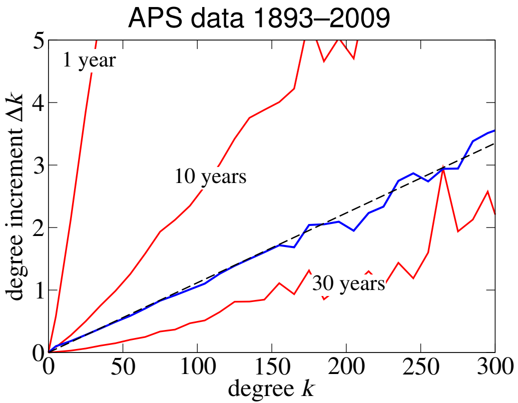
WOMEN AND CHILDREN FIRST

The missing element

The missing element



The missing element



The missing element

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)
  - Independent of the user who makes the link

Do all users react to
item fitness and popularity
in the same way?

Discoverers: A new class of users

- We will show that there are some users who:

  $\boxed{\text{Are often among the first to link with the items that (much) later become very popular}}$

- We call those users discoverers

Discoverers: A new class of users

- We will show that there are some users who:

  > Are often among the first to link with the items that (much) later become very popular

- We call those users discoverers

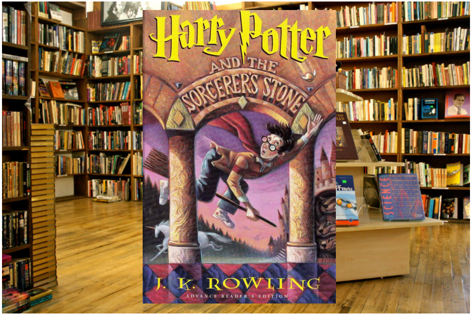
June 26, 1997

Discoverers: A new class of users

- We will show that there are some users who:

  Are often among the first to link with the items that (much) later become very popular

- We call those users discoverers



June 26, 1997

Discoverers: A new class of users

- We will show that there are some users who:

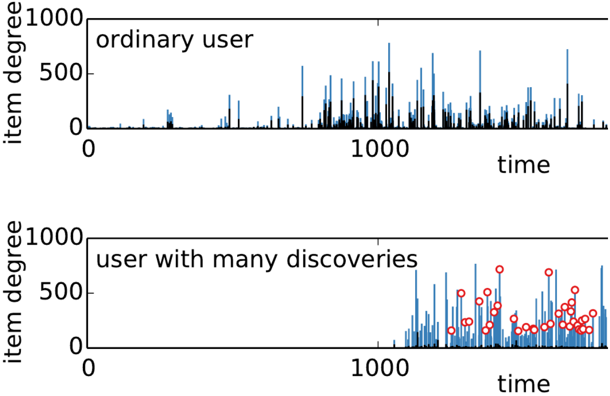
  > Are often among the first to link with the items that (much) later become very popular

- We call those users discoverers

- To find the users who defy popularity, we define:

  > 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).

Discoveries in Amazon data



Black bars: item popularity when collected
Blue bars: final item popularity
Red circles: discoveries

How to quantify the user success

- Use the data to compute:
  - *Number of discoveries* d_i achieved by each user
  - *Number of links* k_i made by each user
- How to assess how unusual is a given user?

How to quantify the user success

- Use the data to compute:
  - *Number of discoveries d_i* achieved by each user
  - *Number of links k_i* made by each user
- How to assess how unusual is a given user?
- Assuming overall discovery probability

$$p_D = \frac{\sum_i d_i}{\sum_i k_i},$$

compute the probability of d_i discoveries or more (user's p-value)

$$P(d \geq d_i | k_i) := P_i^0$$

Top users in the Amazon data

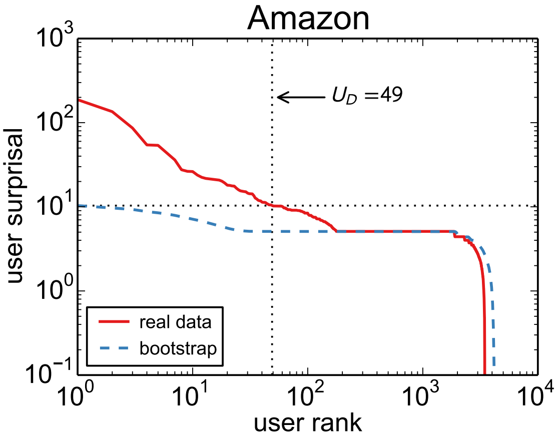
Rank	k_i	d_i	r_i	P_i^0	$s_i = -\ln P_i^0$
1	188	59	51.6	10^{-82}	187.6
2	244	50	33.7	10^{-59}	135.3
3	217	35	26.5	10^{-38}	86.4
4	237	26	18.0	10^{-24}	54.4
5	190	24	20.8	10^{-24}	53.8
6	364	26	11.7	10^{-19}	43.5
7	185	18	16.0	10^{-16}	36.1
8	73	11	24.8	10^{-12}	27.6
9	41	9	36.1	10^{-12}	26.4
10	60	10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

Top users in the Amazon data

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

But: Is this not just luck?

Discoverer or a lucky guy?

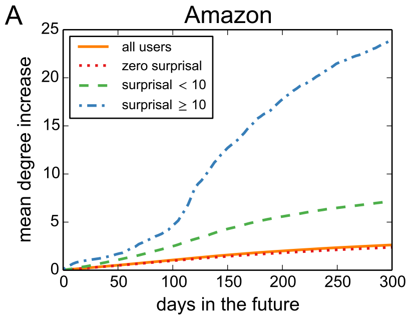


Is this any useful?

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



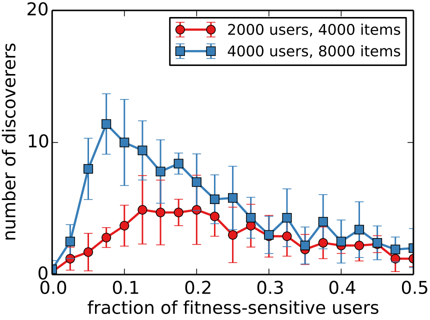
Knowing the discoverers gives us predictive power

A network model

- Network growth model with heterogeneous users
  1. Some users are popularity-driven: $P_i(t) \sim k_i(t) D_R(t)$
  2. Others are fitness-driven: $P_i(t) \sim f_i(t) D_R(t)$

A network model

- Network growth model with heterogeneous users
  1. Some users are popularity-driven: $P_i(t) \sim k_i(t) D_R(t)$
  2. Others are fitness-driven: $P_i(t) \sim f_i(t) D_R(t)$
- The discoverer behavior can be reproduced



Discoverers: conclusions

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

Discoverers: conclusions

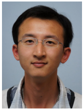
- We find discoverers in almost any information network we look at
- There are still many open questions...
  1. What other influences contribute to the observed discovery patterns? Social status: no. Insider information: partially.
  2. How best to decide who is a discoverer and who is not?
  3. How best to use this information for popularity prediction?
  4. Study the fine struture: maybe someone is a discoverer in sci-fi movies but very ordinary in romantic movies; how to approach this?
  5. How does all this translates to monopartite data?
  6. Connect with the multiple hypothesis testing literature
  7. How to use this knowledge to design better algorithms?

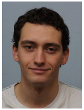
Further related works:

1. M. Medo, G. Cimini, Model-based evaluation of scientific impact indicators, Physical Review E 94, 032312, 2016
2. M. S. Mariani, M. Medo, Y.-C. Zhang, Quantifying the significance of scientific papers through time-balanced network centrality, Submitted to the Journal of Informetrics http://tinyurl.com/rescaled


Giulio Cimini


Stanislao Gualdi


An Zeng


Manuel Mariani

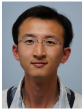

Yi-Cheng Zhang

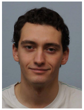
Further related works:

1. M. Medo, G. Cimini, Model-based evaluation of scientific impact indicators, Physical Review E 94, 032312, 2016
2. M. S. Mariani, M. Medo, Y.-C. Zhang, Quantifying the significance of scientific papers through time-balanced network centrality, Submitted to the Journal of Informetrics http://tinyurl.com/rescaled


Giulio Cimini


Stanislao Gualdi


An Zeng


Manuel Mariani


Yi-Cheng Zhang

Thank you for your attention