

Growth of information networks

Matúš Medo, Giulio Cimini, Stanislao Gualdi

Fribourg University, Switzerland

International Workshop on Agent-Based Models
and Complex Techno-Social Systems, ETH Zurich

July 3, 2012



We live in the information age



We live in the information age

Information is created, spreads, fades away



We live in the information age

Information is created, spreads, fades away
this talk

Preferential attachment (PA)

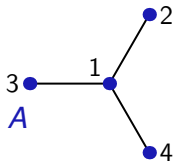
- A classical network model
 - Yule (1925), Simon (1955), Price (1976), Barabási & Albert (1999)
 - Growth of cities, citations of scientific papers, WWW,...

Preferential attachment (PA)

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

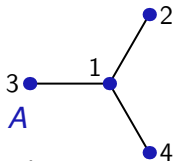
- Nodes and links are added with time

- Probability that a node acquires a new link proportional to its current degree: $P(i, t) \sim k_i(t) + A$

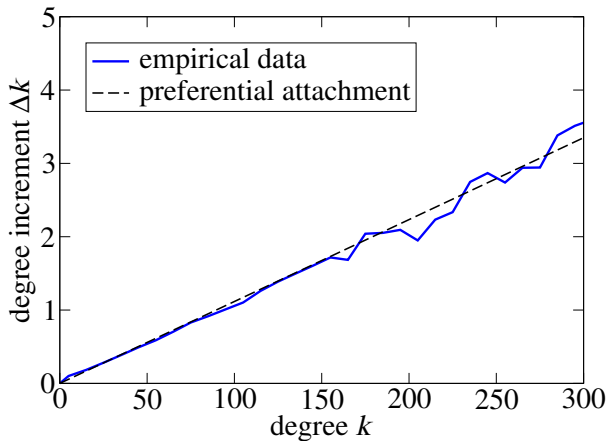


Preferential attachment (PA)

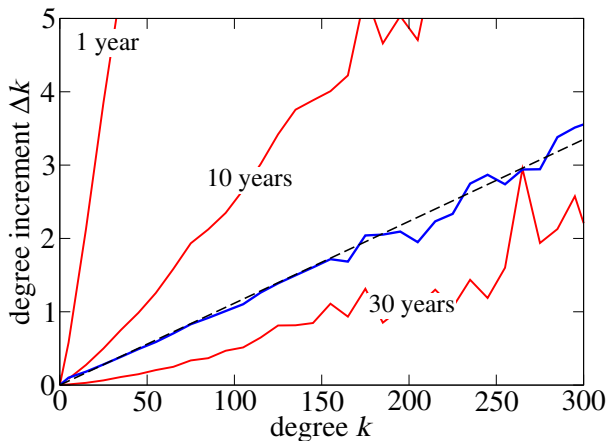
- A classical network model
 - Yule (1925), Simon (1955), Price (1976), Barabási & Albert (1999)
 - Growth of cities, citations of scientific papers, WWW,...
- Nodes and links are added with time
- Probability that a node acquires a new link proportional to its current degree: $P(i, t) \sim k_i(t) + A$
 - Pros: simplicity, resulting power-law degree distribution
 - Cons: simplicity (deviations from the model observed in reality)



PA in scientific citation data



PA in scientific citation data



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

Time decay is fundamental

"All the News That's Fit to Print."

The New York Times.

THE BEASTLER.

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

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 Straus and Wife, and Maj. Butt Aboard.

"HOLE OF DEEP" FOLLOWED

Women and Children Put Safe 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 Lovers Missing.

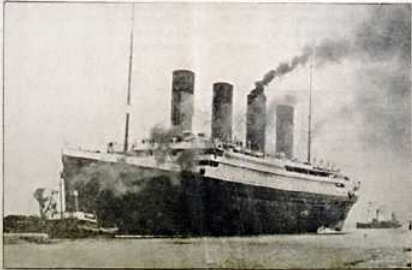
FRANKLIN HOSPITAL ALL SET

Manager of the Life Rafted Thence After Stranahan Dies After She Had Come Aboard.

HEAD OF THE LINE ANKERS

A Boat Seen Being Run Toward Titanic With Four Men to Rescue It There.

The Atlantic, that the Titanic, the largest ship in the world, was sunk at an island and that the ship, the name of which is not yet known, was seen on the 14th of the month.



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 Serpa's Boat to Save Them with the Survivors.

SEA SEARCH FOR OTHERS

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

CLIPPING SENDS THE NEWS

Ship Seen to Head Toward Titanic.

LAST REPORT SAID SHE WAS ON THE 14TH OF THE MONTH.

Two generalizations of the basic PA

- Fitness model (Bianconi & Barabási, 2001):
 - Each node has fitness that influences the attachment probability

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

Two generalizations of the basic PA

- Fitness model (Bianconi & Barabási, 2001):

- Each node has fitness that influences the attachment probability

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

- Aging of sites (Dorogovtsev & Mendes, 2000):

- For a node that appeared at time s , the attachment rate is

$$P(i, t) \sim k_i(t)/(t - \tau_i)^\alpha$$

- They both have their problems...

Outline for the rest

- 1 Formulate a new model
- 2 Present empirical evidence
- 3 Discuss the implications

The model (PRL **107**, 238701, 2011)

- 1 We combine heterogeneous fitness with aging
 - Fitness with aging = relevance

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

- 2 Nodes are not made equal!
 - For example, initial values $R_i(0)$ are random

Solving the 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 model

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

Solving the model

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

⇓

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

Solving the model

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



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

- $R_i(t)$ matters little, it's T_i what's important
- Ω^* can be set to achieve desired $\langle k \rangle$

Degree distributions

- $k_i^F(T_i)$ fluctuates little \implies heterogeneity is needed!

$$\langle k_i^F(T_i) \rangle = \exp(T_i/\Omega^*)$$

Degree distributions

- $k_i^F(T_i)$ fluctuates little \implies heterogeneity is needed!

$$\langle k_i^F(T_i) \rangle = \exp(T_i/\Omega^*)$$

- Some examples:

- 1 $\varrho(T)$ normally distributed \implies log-normal $P(k)$
- 2 $\varrho(T)$ with exponential tail $\implies P(k)$ with a power-law tail
- 3 $\varrho(T) \sim e^{-\alpha T} \implies P(k) \sim k^{-3}$ (exactly as for PA!)

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

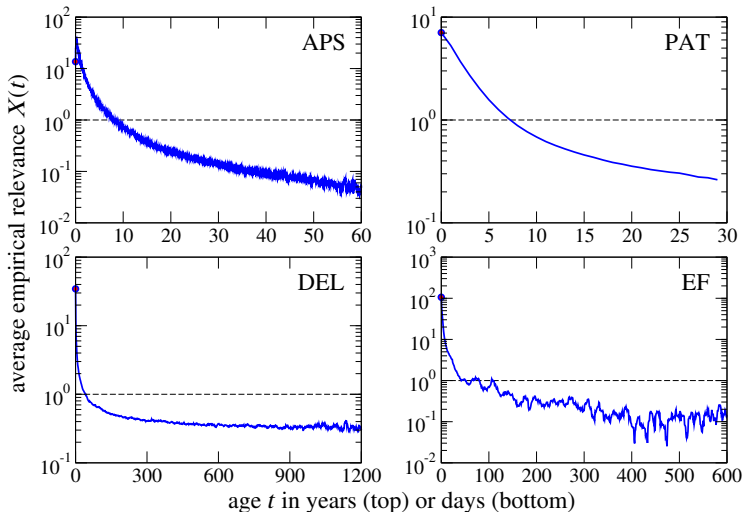
Empirical relevance

- Empirical relevance of paper i at time t : $X_i(t, \Delta t)$

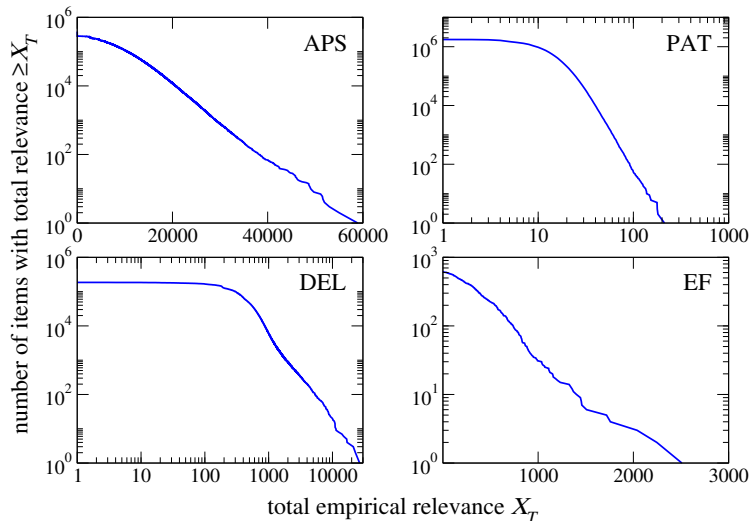
$$X_i(t, \Delta t) := \frac{\text{number of citations received by } i \text{ in } (t, t + \Delta t)}{\text{expected number of citations according to PA}} \sim \frac{\Delta k}{k(t)}$$

- When PA works perfectly, $X_i(t, \Delta t) = 1$

Decay of relevance



Heterogeneity of total relevance



How good model do we have?

- Robust statistical validation needed

How good model do we have?

- Robust statistical validation needed
- Maximum likelihood estimation: maximize $\mathcal{L}(\mathcal{D}|M)$
 - \mathcal{D} is given data (growing network)
 - M is a parametrized network model
 - \mathcal{L} is likelihood of the data for a given model

How good model do we have?

- Robust statistical validation needed
- Maximum likelihood estimation: maximize $\mathcal{L}(\mathcal{D}|M)$
 - \mathcal{D} is given data (growing network)
 - M is a parametrized network model
 - \mathcal{L} is likelihood of the data for a given model
- Problems:
 - **Dimensionality:** number of model parameters proportional to the number of nodes (“high-dimensional statistics”)
 - **Data size:** computation of likelihood is costly
 - **Convergence:** lack of, shallow maximum, validation

Models competing

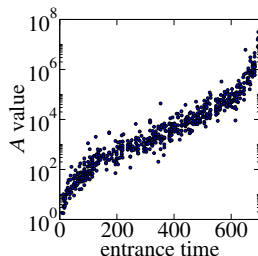
- 0 Links choose target node at random
- 1 PA with an additive term: $P(i, t) \sim k_i(t) + A$
- 2 PA with heterogeneous additive term: $P(i, t) \sim k_i(t) + A_i$
- 3 Eom-Fortunato model: $P(i, t) \sim k_i(t) + A_i(t)$
- 4 PA with relevance: $P(i, t) \sim R_i(t)(k_i(t) + A)$

MLE results - Econophysics Forum data

model	$\max \mathcal{L}(\mathcal{D} M)$	parameters
M_0	-5.55 ± 0.00	0
M_1	-5.52 ± 0.01	1
M_2	-4.44 ± 0.01	N
M_3	-4.00 ± 0.01	$N + 3$
M_4	-3.94 ± 0.01	$N + 4$

MLE results - Econophysics Forum data

model	$\max \mathcal{L}(\mathcal{D} M)$	parameters
M_0	-5.55 ± 0.00	0
M_1	-5.52 ± 0.01	1
M_2	-4.44 ± 0.01	N
M_3	-4.00 ± 0.01	$N + 3$
M_4	-3.94 ± 0.01	$N + 4$



- Pure PA (M_1) almost as bad as the benchmark model M_0
- Heterogeneous additive term (M_3) has little substance
- Difference M_4 vs M_3 seems small but corresponds to \mathcal{D} being 10^{410} -times more likely under M_4 than under M_3

Open problems

- Study clustering coefficient and degree correlations
- Directed nature of the citation network
- Accelerating growth of the network
- Gradual fragmentation into related yet independent fields
- $\Omega(t)$ without a stationary value

Open problems

- Study clustering coefficient and degree correlations
- Directed nature of the citation network
- Accelerating growth of the network
- Gradual fragmentation into related yet independent fields
- $\Omega(t)$ without a stationary value

- Convergence problems for some other datasets
- Knowledge of the dynamics can help select currently most relevant nodes: $\Delta k/k$ matters but. . .

Thank you for your attention!

Questions?