Aging and heterogeneity in the growth of networks

Matúš Medo, Giulio Cimini, Stanislao Gualdi

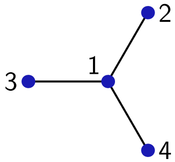
Fribourg University, Switzerland

Conference on Hypernetworks, Network Dynamics and Influence on Networks
December 14, 2011

Growing networks

- Nodes and links are added with time
- Basic model: preferential attachment (PA)
  - Yule (1925), Simon (1955), Price (1976), Barabási & Albert (1999)
  - Growth of cities, citations of scientific papers, WWW,...
  - Probability that a node acquires a new link is assumed proportional to the node's current degree

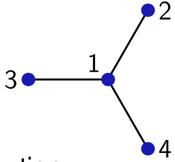
$$P(i, t) \sim k_i(t)$$



Growing networks

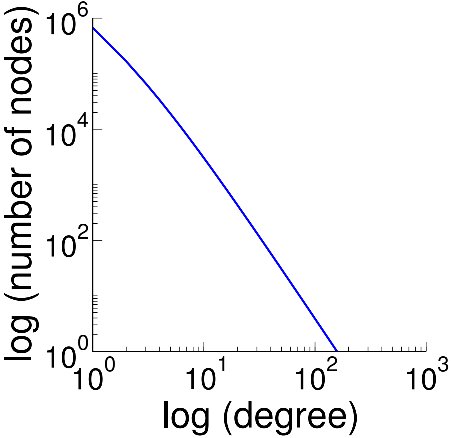
- Nodes and links are added with time
- Basic model: preferential attachment (PA)
  - Yule (1925), Simon (1955), Price (1976), Barabási & Albert (1999)
  - Growth of cities, citations of scientific papers, WWW,...
  - Probability that a node acquires a new link is assumed proportional to the node's current degree

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

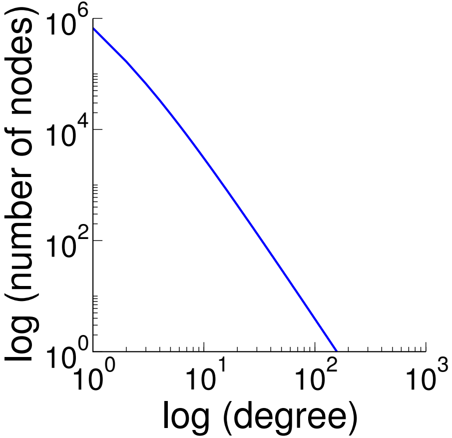


  - Pros: simplicity, resulting power-law degree distribution
  - Cons: simplicity (deviations from the model observed in reality)

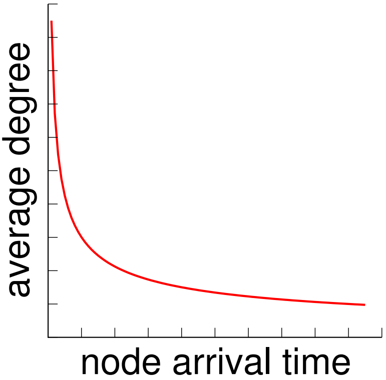
Pros/cons



Pros/cons



log (number of nodes)
10^6
10^4
10^2
10^0
10^0
10^1
10^2
10^3
log (degree)



average degree
node arrival time

Cons continued

- Many distributions claimed in the literature to be power laws fail in rigorous statistical tests (Clauset, Shalizi, Newman, 2009)
- Citation data shows patterns different from PA (Redner, 2005)
- No correlation between the age of a site and its number of incoming links in the WWW (Adamic & Huberman, 2000)
- A first-mover advantage in scientific citations exists but notable exceptions are present (Newman, 2009):
  "(There is) a hopeful sign that we as scientists do pay at least some attention to good papers that come along later"

Two generalizations of the basic PA

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

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

  - Fitness distribution with unbounded support $\implies$ link condensation

Two generalizations of the basic PA

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

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

  - Fitness distribution with unbounded support $\implies$ link condensation
- Aging of sites (Dorogovtsev & Mendes, 2000):
  - For a node that appeared at time s, the attachment rate is

$$P(i,t) \sim k_i(t)/(t-s)^{\alpha}$$

  - Scale-free $P(k)$ is observed only for very slow decay ($\alpha < 1$)

Outline for the rest

1. Formulate a new model
2. Present empirical evidence
3. Solve the model
4. Discuss the implications

New 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. Important point: not all nodes are equal
   - For example, initial values $R_i(0)$ are random

New 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. Important point: not all nodes are equal
   - For example, initial values $R_i(0)$ are random

But is this really relevant?

Empirical evidence

- Citation data provided by the American Physical Society
  - 450'084 papers published by the APS from 1893 to 2009
  - 4'691'938 citations within the APS journals
- In-degree distribution:
  - $\alpha = 2.29 \pm 0.01$, $x_{\mathrm{min}} = 50$
  - Statistical significance only for $x_{\mathrm{min}} \gtrsim 150$
  - Log-normal distribution does not fit the data better

Empirical evidence

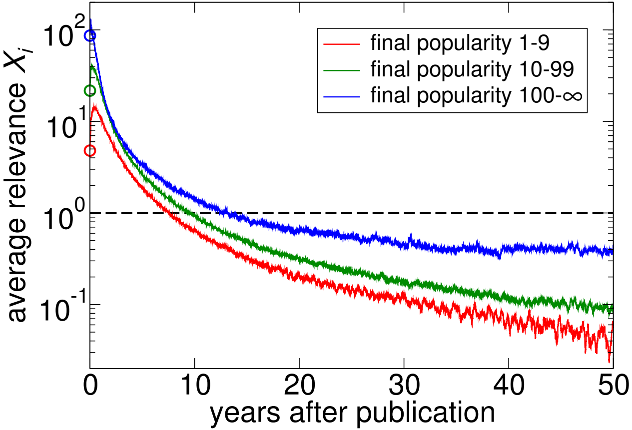
- Citation data provided by the American Physical Society
  - 450'084 papers published by the APS from 1893 to 2009
  - 4'691'938 citations within the APS journals
- In-degree distribution:
  - $\alpha = 2.29 \pm 0.01$, $x_{\text{min}} = 50$
  - Statistical significance only for $x_{\text{min}} \gtrsim 150$
  - Log-normal distribution does not fit the data better
- Empirical relevance of paper i at time t: $X_i(t, \Delta t)$

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

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

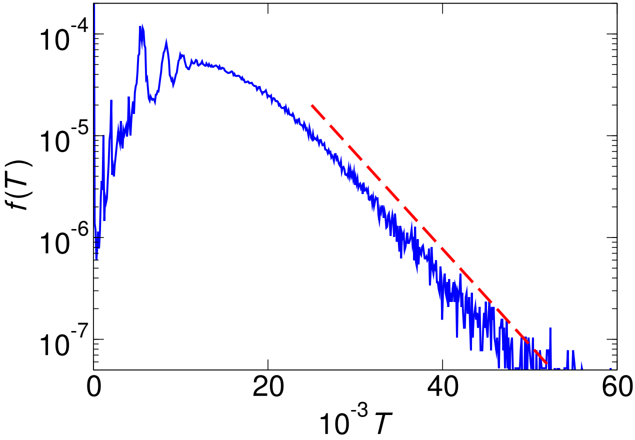
Decay of relevance in the APS data

time window $\Delta t = 91$ days



Heterogeneity of total relevance in the APS data

$$T_i := \sum_t X_i(t)$$

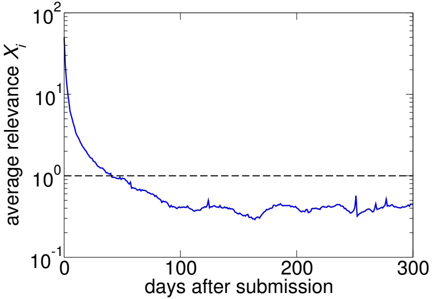


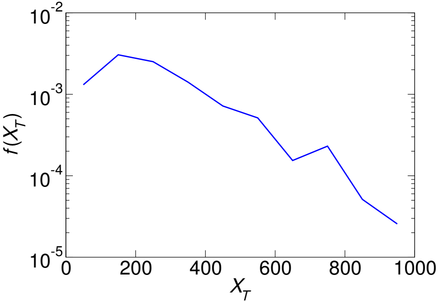
The case of the Econophysics Forum

- A site for researchers in Econophysics
  - `www.unifr.ch/econophysics`
- 390 papers submitted from July 2010 until August 2011
  - 19 320 downloads (50 per paper) analyzed with $\Delta t = 30$ days

The case of the Econophysics Forum

- A site for researchers in Econophysics
  - `www.unifr.ch/econophysics`
- 390 papers submitted from July 2010 until August 2011
  - 19 320 downloads (50 per paper) analyzed with $\Delta t = 30$ days





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

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

Solving the model

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

$$\Downarrow$$

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

Solving the model

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

$$\Downarrow$$

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

- The form of $R(t)$ matters little: it's T what's important
- Ω^* determined by self-consistency: the average degree is two

$$\int \varrho(T)\,\mathrm{e}^{T/\Omega^*}\mathrm{d}T = 2 \qquad (\varrho(T) \implies \Omega^*)$$

Degree distributions

- When T_i is given, k_i^F fluctuates little
- To model real networks, heterogeneous T is needed

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

Degree distributions

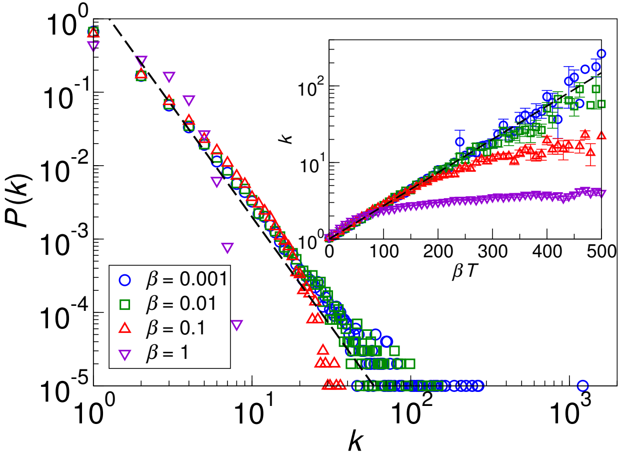
- When T_i is given, k_i^F fluctuates little
- To model real networks, heterogeneous T is needed

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

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

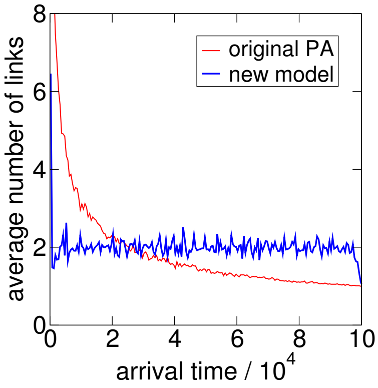
Numerical results

$R_i(t) = R_i(0)e^{-\beta(t-t_i)}$, $R_i(0)$ exponentially distributed



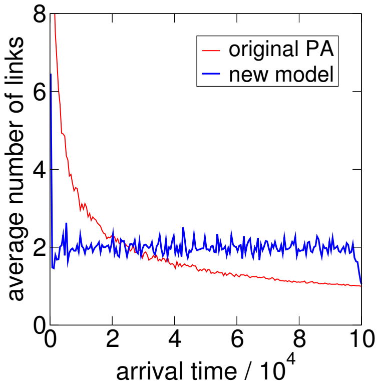
Time bias removed

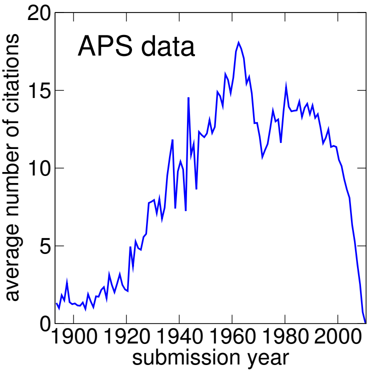
Average degree vs age



Time bias removed

Average degree vs age





Summary

- Aging and heterogeneity combined in a new model
- Solves the time bias problem of PA
- Evidence from citation data and website users
- Should be applicable to many information networks

Open questions

- 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 questions

- 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

- Why $\varrho(T)$ for citation data shows an exponential tail?
- What about other systems where PA is at work?

Challenges

- Mitzenmacher (2005): types of results when studying power laws
  1. *Observe*: Gather data and demonstrate a power law fit
  2. *Interpret*: Explain the significance of the power law behavior
  3. *Model*: Propose an underlying model that explains it
  4. *Validate*: Find data to validate/modify the model
  5. *Control*: Use the understanding from the model to control, modify, and improve the system behavior

Challenges

- Mitzenmacher (2005): types of results when studying power laws
  1. *Observe*: Gather data and demonstrate a power law fit
  2. *Interpret*: Explain the significance of the power law behavior
  3. *Model*: Propose an underlying model that explains it
  4. *Validate*: Find data to validate/modify the model
  5. *Control*: Use the understanding from the model to control, modify, and improve the system behavior
- Ad 4: Maximum Likelihood Estimation can help fit individual relevance values
- Ad 5: Knowledge of the dynamics can help select the (currently) most relevant nodes

Thank you for your attention