Temporal patterns in information and social systems

Matúš Medo

University of Fribourg, Switzerland

COST Workshop
Quantifying scientific impact: networks, measures, insights?

12-13 February, 2015, Zurich

Outline

1. Growing networks with fitness and aging
2. Temporal bias of PageRank
3. Leaders and followers and the consequences

The common theme

Temporal patterns and the role of time in information and social systems.

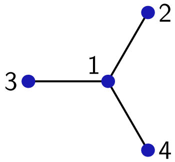
Preferential attachment (PA)

- 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

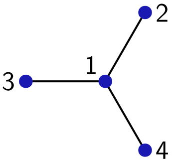
$$P(i, t) \sim k_i(t)$$



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

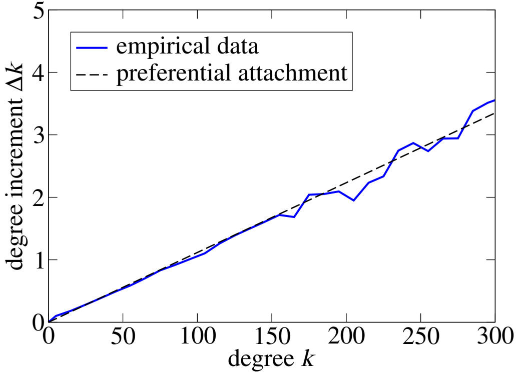
$$P(i,t) \sim k_i(t)$$



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

PA in scientific citation data

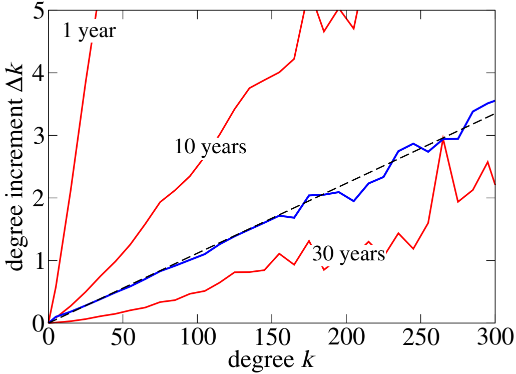
Journals of the American Physical Society from 1893 to 2009:



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.

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

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 (PRE 89, 032801, 2014)

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

- Point to note:

Paper popularity grows exponentially with fitness (quality)

$\Downarrow$

Fitness depends logarithmically on popularity

Two forms of aging in information networks

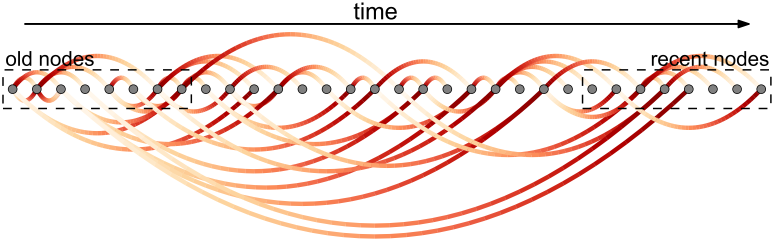
- The decay of relevance: $D_R(t)$
  - Node relevance influences the in-coming links

Two forms of aging in information networks

- The decay of relevance: $D_R(t)$
  - Node relevance influences the in-coming links
- The decay of activity: $D_A(t)$
  - Nodes activity influences the out-going links
  - Activity decays in time (mostly)

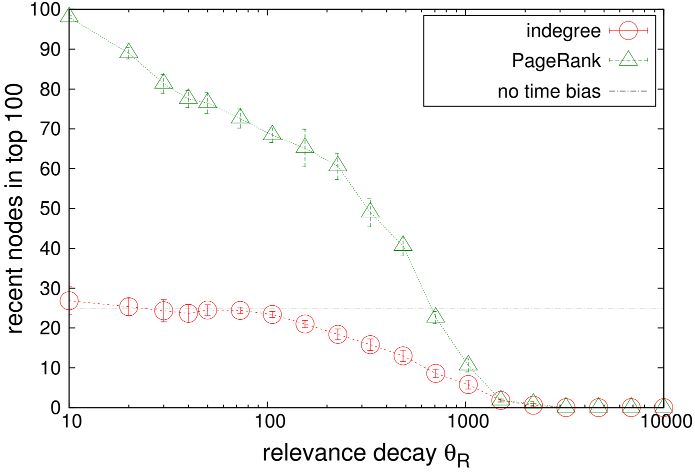
Two forms of aging in information networks

- The decay of relevance: $D_R(t)$
  - Node relevance influences the in-coming links
- The decay of activity: $D_A(t)$
  - Nodes activity influences the out-going links
  - Activity decays in time (mostly)

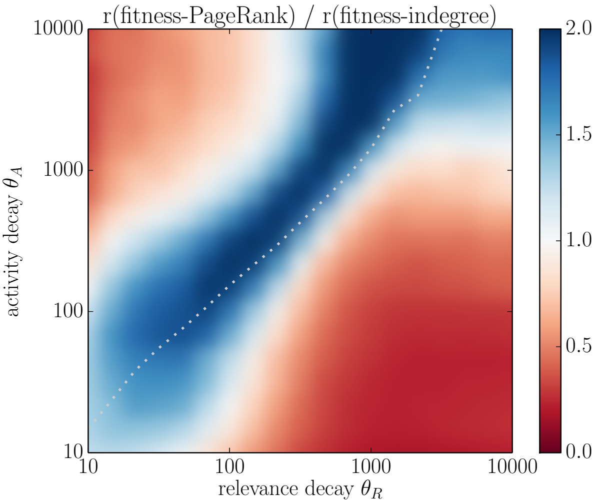


A growing network with a quick decay of attractiveness and no decay of activity

The biases of PageRank



The biases of PageRank



The biases of PageRank

Implications

In citation data, the time scales of relevance and activity decay are very different ($\Theta_A = 0$ because outgoing links are created only upon arrival).
PageRank (and its variants) is nevertheless commonly applied on citation data. One should think twice!

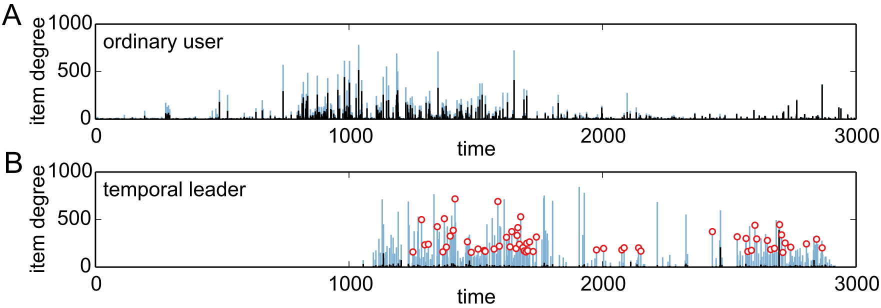
The case for leaders in social systems

- Bipartite user-item data (*who* bought *what* at Amazon.com)
  - Similar behavior in monopartite social data (user-user)
- Most users are driven by item popularity (*followers*)
- Some users are driven by item fitness (*leaders*)

The case for leaders in social systems

- Bipartite user-item data (*who* bought *what* at Amazon.com)
  - Similar behavior in monopartite social data (user-user)
- Most users are driven by item popularity (*followers*)
- Some users are driven by item fitness (*leaders*)
- 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

Leaders in Amazon data



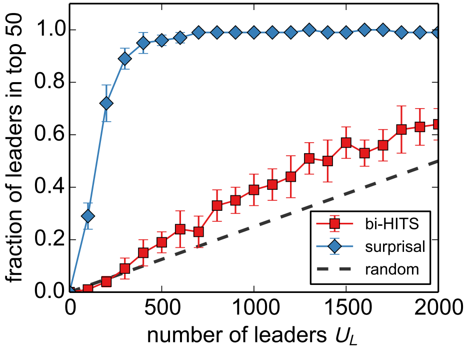
Black bars: popularity of collected items when they are collected.
Blue bars: final popularity of collected items.
Red circles: discoveries.

The game changer

- Network growth model with to rules reproduces the real data patterns
  1. Followers choose items driven by $k_i(t)D_R(t)$
  2. Leaders choose items driven by $f_i(t)D_R(t)$

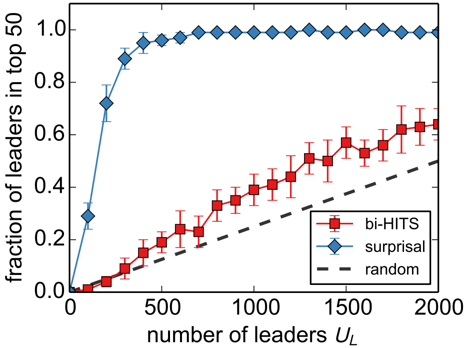
The game changer

- Network growth model with to rules reproduces the real data patterns
  1. Followers choose items driven by $k_i(t)D_R(t)$
  2. Leaders choose items driven by $f_i(t)D_R(t)$
- Model data poses a puzzle to classical ranking algorithms



The game changer

- Network growth model with to rules reproduces the real data patterns
  1. Followers choose items driven by $k_i(t)D_R(t)$
  2. Leaders choose items driven by $f_i(t)D_R(t)$
- Model data poses a puzzle to classical ranking algorithms



Reason: Insightful choices of the leaders are copied by the followers. All users ultimately collect items of the same fitness and an algorithm acting on a static data snapshot cannot distinguish them.
Solution: Algorithms that take time into account adequately.

Coming back to the guiding questions:

“Are the implicit assumptions of centrality measures justified in scientometrics?”
“It doesn’t seem to be the case!”

“Are altmetrics shallow?”
“Build on the community structure of science!”
(PLoS ONE 9, e112022, 2014)

Thank you for your attention