Citation networks as a window to science: a case study

Remco van der Hofstad

Symposium facets of Data Analytics,
April 21, 2017, TU/e

Joint work with
- Alessandro Garavaglia
- Gerhard Woeginger
Citations

Citation counts contain important information, yet are hard to interpret:

- Depend extremely on age scientists;
- Highly field dependent (even differences within small subfields);
- Many good papers with few, and bad papers with many citations;
- Metrics (such as h-index and journal impact factors) have obvious limitations.

Let’s make science metrics more scientific.
Look at the data!

- Investigate citation network dynamics:
  - How many citations do papers receive?
  - What is variability in citation counts?
  - How long does it take for papers to receive citations?
  - When is your paper forgotten?

- Restrict to homogeneous domains of science:
  - Probability and statistics
  - Electrical engineering
  - Biomedical technology

On basis of Web of Science data [not good for CS]:
  - 40 M papers with 500 M citations starting in 1980.
  - Courtesy of CWTS Leiden (Ludo Waltman)
Exponential growth of number of publications. Already observed by Derek De Solla Price in his 1963 book ‘Little Science, big science’
Citations of papers

Extreme variability in citation distributions: Power laws?
Dynamic power laws

Power laws change over time: 
Citation distribution of papers from 1984 in Probability and Statistics
Dynamic power laws

Power laws change over time:
Citation distribution of papers from 1984 in Biomedical Technology
Evolution of citations over time:
Random sample of 20 papers from 1980
Evolution of citations

Average citation increment over a 20-years time window for papers published in different years.
Aging of citations

Distribution of age of cited papers for different citing years:
Probability and Statistics
Aging of citations

Distribution of age of cited papers for different citing years:
Biomedical Engineering
Almost linear growth citations

Average number of citations received by papers published in 1984 in 1993, 2006 and 2013 according to total citations up to same year.
Conclusions

- Number of papers grows almost exponentially;
- Citations per paper vary tremendously;
- Citation counts follow approximate power-law distribution, with exponent changing over time;
- Papers stop receiving citations after (random) time;
- Age of cited papers looks roughly log-normal;
- Reasonable prediction that citations grow almost linearly in time given past.
Modeling networks

Use random graphs to model uncertainty in formation connections between elements.

- **Static models:**
  Graph has fixed number of elements:
  - **Erdős-Rényi random graph** and **configuration model**.

- **Dynamic models:**
  Graph has evolving number of elements:
  - **Preferential attachment model**

Due to highly dynamic nature of citation networks, focus on dynamic models.
At time $n$, single vertex is added with $m$ edges emanating from it. Probability that edge connects to $i$th vertex is proportional to

$$D_i(n - 1) + \delta,$$

where $D_i(n)$ is degree vertex $i$ at time $n$, $\delta > -m$ is parameter.

Yields power-law degree sequence with exponent

$$\tau = 3 + \delta/m > 2.$$

$m = 2, \delta = 0, \tau = 3, n = 10^6$
Preferential attachment

- Preferential attachment models (PAMs) grow **linearly** in time:
  - Embed in continuous time, where growth becomes exponential.
  - Idea fruitful also for regular PAMs: Rudas, Tóth and Valko (2007):
    - availability of powerful tools of continuous-time branching processes.

- **Old-get-richer** phenomenon leading in PAMs:
  - Introduce random fitness for each vertex in graph.

- Vertices keep on receiving citations in PAMs:
  - Introduce aging effect.
Model

- Preferential attachment model in continuous time where rate of growth at time $t$ of links to vertex $v$ that is born at time $s$ are given by

$$\eta_v(D_v(t) + \delta)g(t - s),$$

where

- $\eta_v$ is fitness of vertex $v$:

  Citation counts become highly variable;

- $g$ is (integrable) aging function:

  Vertices receive finite number of citations in lifetime;

- $D_v(t)$ is degree of vertex $v$ at time $t$:

  Increments of citation counts roughly linear;

- $\delta$ is parameter allowing for fine tuning.
Examples of degree distributions with/without aging and fitness
Dynamic power law

Simulation of dynamic power law
Results

Reasonable **qualitative** comparison model/data:

- Exponential growth;
- Integrable aging;
- Highly variable fitnesses.

Rigorous results in tree case:

- Exponential growth is **typical behavior**;
- Power laws with aging and fitness **only** when fitness has **at most** exponential tail.
Challenges

(A) Make model comparison quantitative:
- Initialization difficult due to lack of data before 1980;
- Estimate model parameters from data;
- Estimating fitness might lead to citation prediction;

(B) Extend rigorous results to non-tree setting:
- Possibility through collapsing multiple tree vertices;
- Difficulty in making fitnesses agree.

(C) Extend model to include scientists:
- Hard to deal with varying research teams;
- Relate fitness to attributes, such as authors and journals.

Back to scientists we started out with!
Aimed at graduate students in math. Also informal explanations of random graphs for networks and basic results. Non-mathematicians can ignore proofs...

www.networkpages.nl