Inferring the causal effect of journals on citations

V.A. Traag

Centre for Science and Technology Studies (CWTS)
Leiden University, the Netherlands

Complexity Science Hub Vienna, 5 February 2021
High-impact journals are more highly cited.

Why?

1. High-impact journals publish more “citable” articles?
2. Increase citations because articles appeared in high-impact journals?
High-impact journals are more highly cited.

Why?

1. High-impact journals publish more "citable" articles?
2. Increase citations because articles appeared in high-impact journal?
High-impact journals are more highly cited.

Why?

1. High-impact journals publish more “citable” articles?
2. Increase citations because articles appeared in high-impact journal?
Causal model

Value

Journal → Citations
Causal model

Value

Journal  Citations

Evaluation
Causal model

Value

Journal

Citations

Evaluation
Causal model

Value

Journal → Citation

Evaluation
Causal model

Value → Pre-print citations

Journal → Value → Citations
Causal model

Latent citation rate

Method
Novelty
Reputation

Journal
Citations

Pre-print citations
Causal model

Latent citation rate

Method  Novelty  . . .  Reputation

Pre-print duration

Journal  Citations

Pre-print citations
Citations $c_i(t)$ that publication $i$ receives at time $t$

$$c_i(t) \sim \text{Poisson} \left[ \lambda_i(t) \times f_i(t) \times (m + C_i(t - 1)) \right]$$

---

Wang, D, Song, C & Barabási, AL. *Science* **342**, 127–132 (2013).
Citations $c_i(t)$ that publication $i$ receives at time $t$

$$c_i(t) \sim \text{Poisson} \left[ \lambda_i(t) \times f_i(t) \times (m + C_i(t - 1)) \right]$$

- Effective citation rate

---

Wang, D, Song, C & Barabási, AL. *Science* **342**, 127–132 (2013).
Model

Citations $c_i(t)$ that publication $i$ receives at time $t$

$$c_i(t) \sim \text{Poisson} \left[ \lambda_i(t) \times f_i(t) \times (m + C_i(t - 1)) \right]$$

- Effective citation rate
- Temporal decay

---

Wang, D, Song, C & Barabási, AL. *Science* **342**, 127–132 (2013).
Citations $c_i(t)$ that publication $i$ receives at time $t$

$$c_i(t) \sim \text{Poisson} \left[ \lambda_i(t) \times f_i(t) \times (m + C_i(t-1)) \right]$$

- Effective citation rate
- Temporal decay
- Cumulative advantage

---

Wang, D, Song, C & Barabási, AL. *Science* **342**, 127–132 (2013).
Model

Citations $c_i(t)$ that publication $i$ receives at time $t$

$$c_i(t) \sim \text{Poisson} \left[ \lambda_i(t) \times f_i(t) \times (m + C_i(t - 1)) \right]$$

- Effective citation rate
- Temporal decay
- Cumulative advantage

Effective citation rate

Wang, D, Song, C & Barabási, AL. *Science* **342**, 127–132 (2013).
Model

Citations $c_i(t)$ that publication $i$ receives at time $t$

$$c_i(t) \sim \text{Poisson} \left[ \lambda_i(t) \times f_i(t) \times (m + C_i(t - 1)) \right]$$

- Effective citation rate
- Temporal decay
- Cumulative advantage

Effective citation rate

Pre-publication citation rate

$$\lambda_i(t) = \phi_i$$

Wang, D, Song, C & Barabási, AL. *Science* **342**, 127–132 (2013).
Model

Citations $c_i(t)$ that publication $i$ receives at time $t$

$$c_i(t) \sim \text{Poisson} \left[ \lambda_i(t) \times f_i(t) \times (m + C_i(t - 1)) \right]$$

- Effective citation rate
- Temporal decay
- Cumulative advantage

Effective citation rate

Pre-publication citation rate $\lambda_i(t) = \phi_i$ ← Latent citation rate

Wang, D, Song, C & Barabási, AL. *Science* 342, 127–132 (2013).
Citations $c_i(t)$ that publication $i$ receives at time $t$

$$c_i(t) \sim \text{Poisson} \left[ \lambda_i(t) \times f_i(t) \times (m + C_i(t-1)) \right]$$

- Effective citation rate
- Temporal decay
- Cumulative advantage

Effective citation rate

- Pre-publication citation rate $\lambda_i(t) = \phi_i$
- Post-publication citation rate $\lambda_i(t) = \phi_i \times \theta_j$

Wang, D, Song, C & Barabási, AL. *Science* 342, 127–132 (2013).
Model

Citations $c_i(t)$ that publication $i$ receives at time $t$

$$c_i(t) \sim \text{Poisson} \left[ \lambda_i(t) \times f_i(t) \times (m + C_i(t - 1)) \right]$$

- Effective citation rate
- Temporal decay
- Cumulative advantage

Effective citation rate

Pre-publication citation rate
$$\lambda_i(t) = \phi_i$$

Latent citation rate

Post-publication citation rate
$$\lambda_i(t) = \phi_i \times \theta_{J_i}$$

Journal citation multiplier

---

Wang, D, Song, C & Barabási, AL. *Science* 342, 127–132 (2013).
Citation dynamics (astro-ph/0405353)
Hierarchical Bayesian model – Latent citation rate prior

Each article $i$ in journal $j$, $\phi_i \sim \text{LogNormal}(\Phi_j, \varepsilon_j)$. 
Hierarchical Bayesian model – Temporal decay prior

Decay $f_i(t)$ is exponential, with inverse rate $\beta_i$ (prior average $\approx$ 3 years).

![Probability density](A)

![Temporal decay](B)
We assume $\theta_j$ is centered around 1 (i.e. no increase/decrease).
arXiv

- Collect 727,186 preprints with DOI from arXiv (≈ 54% of all arXiv preprints).
- Get publication date from CrossRef using DOI.
  - 521,512 preprints put on arXiv before published.
- Use arXiv subject for scientific fields.
Data

arXiv

- Collect 727,186 preprints with DOI from arXiv (≈ 54% of all arXiv preprints).
- Get publication date from CrossRef using DOI.
  - 521,512 preprints put on arXiv before published.
- Use arXiv subject for scientific fields.

Scopus

- Use Scopus to get citations to preprint (using arXiv identifier) and published version.
- Pre-publication citations: before publication date (156,528).
- Post-publication citations: after publication date (15,939,887).
- Calculate journal impact from Scopus.
  - Average citations in 5 years for articles & reviews.
Citations and impact

A

B

C

D

In Figure A, we see a scatter plot with post-publication citations on the y-axis and pre-publication citations on the x-axis. The distribution suggests a strong correlation between pre-publication and post-publication citations.

In Figure B, the x-axis represents journal impact and the y-axis shows pre-publication citations. The data indicates that higher journal impact is associated with a greater number of pre-publication citations.

Figure C presents a scatter plot with preprint duration on the x-axis and pre-publication citations on the y-axis. It appears that longer preprint durations correlate with more citations.

Finally, Figure D plots post/pre citations against journal impact, showing a trend where higher journal impact is linked to a higher ratio of post-to-pre citations.

These visualizations provide insights into the dynamics of academic research, with implications for citation metrics and impact factors.
Model fit - article citations

A

B

![Graph showing model fit with article citations]

- 95% bounds
- Median

Pre-publication citations

Post-publication citations
Why are high-impact journals cited more?

1. High-impact journals publish articles with higher latent citation rates.
2. High-impact journals increase citations more (cumulative advantage).

Considerations

- Causal effect presumably related to readership/circulation.
- Self-perpetuating journal hierarchy (high-impact because it has high-impact).
- Journal may be useful conduit of information for citations.
Conjectured causal model

Method → Novelty → Journal → Latent citation rate → Citations → Evaluation
Method → Novelty → Journal → Relevance → Reputation
Method → Evaluation
Method → Novelty
Method → Relevance
Method → Reputation
Novelty → Journal
Novelty → Citations
Novelty → Evaluation
Novelty → Relevance
Novelty → Reputation
Relevance → Journal
Relevance → Citations
Relevance → Evaluation
Relevance → Reputation
Reputation → Journal
Reputation → Citations
Reputation → Evaluation
Reputation → Relevance
Reputation → Novelty
Thank you!

Suggestions? Comments? Questions?

✉️ v.a.traag@cwts.leidenuniv.nl  
🐦 @vtraag