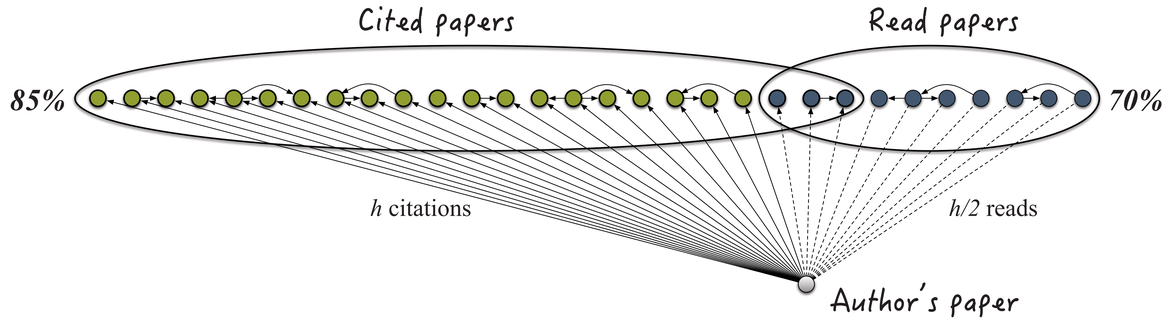
Who reads and who cites?

Unveiling author citation dynamics by modeling citation networks

Lovro Šubelj*, Slavko Žitnik & Marko Bajec
University of Ljubljana, Faculty of Computer and Information Science
*Corresponding author: lovro.subelj@fri.uni-lj.si



Background

Over *10* years ago, Simkin & Roychowdhury [1] showed that around ***80%* cited papers are never read** but merely copied from the bibliographies of other papers! Their study was based on the misprints in bibliographies.

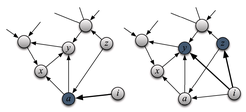
Methodology

We derive **realistic graph model of citation networks** [2] that mimics an author including references into bibliography of a paper. Modelling **author citation dynamics** allows for different applications in bibliometrics & scientometrics!

Methods & Data

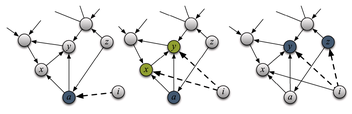
Forest fire model

A new node *i* chooses an ambassador *a* and links to it (solid lines). Next, some of its neighbors are taken as ambassadors by following in- and out-links with probabilities p_f and p_b (*y* and *z*). [4]



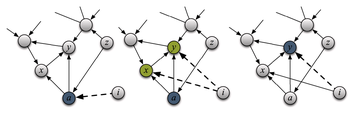
Citation model (our)

A new node *i* links to *a* with probability q_a (dashed lines) and also to its neighbors with probability q_l by following out-links (*x* and *y*). Next, some of its neighbors are taken as ambassadors by following in- and out-links with probabilities p_f and p_b (*y* and *z*). [2,3]



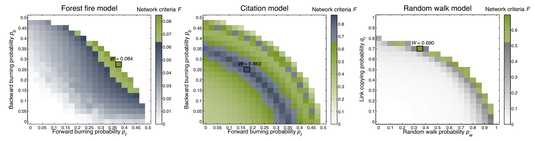
Random walk model

A new node *i* traverses the graph in a random walk fashion by following a single out-link with probability p_w (linking dynamics are as above). [5]



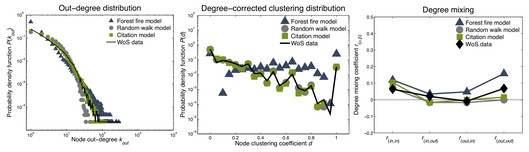
Parameter estimation

Model parameters are estimated by stochastic gradient descent based on network criteria function F that combines *10* standard graph metrics.



Graph structure

Citation model well reproduces graph structure of citation networks. *Forest fire* overestimates the clustering, while *Random walk* underestimates the out-degrees.



Bibliometrics & scientometrics

 # Cited papers:

$$\langle k_{out} \rangle$$

 # Read papers:

$$s = \left(1 - \frac{p_f}{1-p_f} - \frac{p_b}{1-p_b}\right)^{-1}$$

% Cited paper is read:

$$1 - \frac{sq_l}{(1-q_l)\langle k_{out} \rangle}$$

% Read paper is cited:

$$q_a$$

For details and applications in bibliometrics & scientometrics see [2,3].

Web of Science data

The analyses are based on over *60* years of *Web of Science* data including *750,996* journal papers and *1,668,168* citations. Particularly, we consider WoS categories *Information & Library Science* (ILS), *Computer Science, Theory & Methods* (TM), *Software Engineering* (SE), *Artificial Intelligence* (AI), and *Cybernetics* (CY).

Highlights

In most cases, the authors **who cite a paper do not read the paper** (& vice-versa).
Throughout the years, the authors began to **read less and cite more papers**.
If a paper has been **cited *h*-times**, it has been **read around *h/2*-times**.
Author reading dynamics are consistent across the fields.

Results & Discussion

Scientific field comparison

The number of papers cited by a published paper depends on the field of study, however, **the number of papers read by the authors is independent of the field!** The percentage of citations merely copied from other papers is around ***80-85%***, while the probability of citing a read paper is around ***30-45%***.

	Paper citation		Paper study		Paper discovery by		
Data	# Cite	% Copy	# Read	% Cite	% Citation	% Service	% Other
ILS	3.98	86.1%	2.14	27.9%	29.2%	41.0%	29.8%
TM	2.93	79.7%	1.47	45.2%	74.7%	0.5%	24.9%
AI	4.52	87.3%	1.47	40.9%	25.8%	47.6%	26.6%
SE	2.78	81.5%	1.58	36.4%	68.8%	2.0%	29.2%
CY	2.18	69.6%	1.59	43.2%	24.5%	37.8%	37.6%

Temporal bibliometric analysis

Citation dynamics have changed notably over the years, whereas authors **read less and cite more papers!** In 1970s, more papers were read than cited, while nowadays **only a single paper is read for every two cited.** The percentage of papers discovered through citations has remained roughly the same, while the percentage of papers discovered through online services has increased with the growth of the Internet in the 1990s.

	Paper citation		Paper study		Paper discovery by		
Period	# Cite	% Copy	# Read	% Cite	% Citation	% Service	% Other
1945–2013	3.98	86.1%	2.14	27.9%	29.2%	41.0%	29.8%
1970–1980	2.23	52.1%	3.39	33.5%	41.4%	0.0%	58.5%
1980–1990	2.62	65.1%	2.96	33.0%	48.3%	1.1%	50.6%
1990–2000	3.42	81.6%	2.38	29.0%	40.3%	23.2%	36.5%
2000–2010	5.06	83.6%	2.90	32.2%	40.7%	27.5%	31.7%

[1] M. V. Simkin & V. P. Roychowdhury. Read before you cite! *Compl. Syst.*, 14:269–274, 2003.
[2] L. Šubelj, D. Fiala, S. Žitnik & M. Bajec. Modeling citation network topology, *in preparation*, 2014.
[3] L. Šubelj & M. Bajec. Model of complex networks based on citation dynamics. In *Proceedings of the WWW Workshop on Large Scale Network Analysis*, p. 527, 2013.
[4] J. Leskovec, J. Kleinberg & C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1):1–41, 2007.
[5] A. Vazquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Phys. Rev. E*, 67(5):056104, 2003.