Collective Attention towards Scientists and Research Topics

Claudia Wagner
GESIS - Leibniz Institute for the Social Sciences and U. of Koblenz-Landau
Cologne/Koblenz, Germany
claudia.wagner@gesis.org

Olga Zagovora
GESIS - Leibniz Institute for the Social Sciences
Cologne, Germany
olga.zagovora@gesis.org

Tatiana Sennikova
U. of Koblenz-Landau
Koblenz, Germany
tsennikova@uni-koblenz.de

Fariba Karimi
GESIS - Leibniz Institute for the Social Sciences
Cologne, Germany
fariba.karimi@gesis.org

ABSTRACT

Emergent patterns of collective attention towards scientists and their research may function as a proxy for scientific impact which traditionally is assessed via committees that award prizes to scientists. Therefore it is crucial to understand the relationships between scientific impact and online demand and supply for information about scientists and their work. In this paper, we compare the temporal pattern of information supply (article creations) and information demand (article views) on Wikipedia for two groups of scientists: scientists who received one of the most prestigious awards in their field and influential scientists from the same field who did not receive an award.

Our research highlights that awards function as external shocks which increase supply and demand for information about scientists, but hardly affect information supply and demand for their research topics. Further, we find interesting differences in the temporal ordering of information supply between the two groups: (i) award-winners have a higher probability that interest in them precedes interest in their work; (ii) for award winners interest in articles about them and their work is temporally more clustered than for non-awarded scientists.

KEYWORDS

altmetrics; social-media-metrics; online attention; science of science; Wikipedia

ACM Reference Format:
Claudia Wagner, Olga Zagovora, Tatiana Sennikova, and Fariba Karimi. 2018. Collective Attention towards Scientists and Research Topics. In *WebSci '18: 10th ACM Conference on Web Science, May 27–30, 2018, Amsterdam, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3201064.3201097

1 INTRODUCTION

The temporal dynamics of online information supply and demand [5] for research topics and scientists may reveal information about their impact. For example, if a scientist innovates a new research topic, the interest of the general public in the topic will be most likely connected with the interest in the scientists (or the other way around). Therefore the interest will be temporally clustered. If interest in the scientist increases, the topic will probably also gain interest or vice versa. Conversely, if a scientist's works on a research topic had attracted attention from the general public long before anyone was interested in the scientist, then the interest in the research topic was not driven by the scientist since the temporal order is a necessary (but not a sufficient) condition for causality.

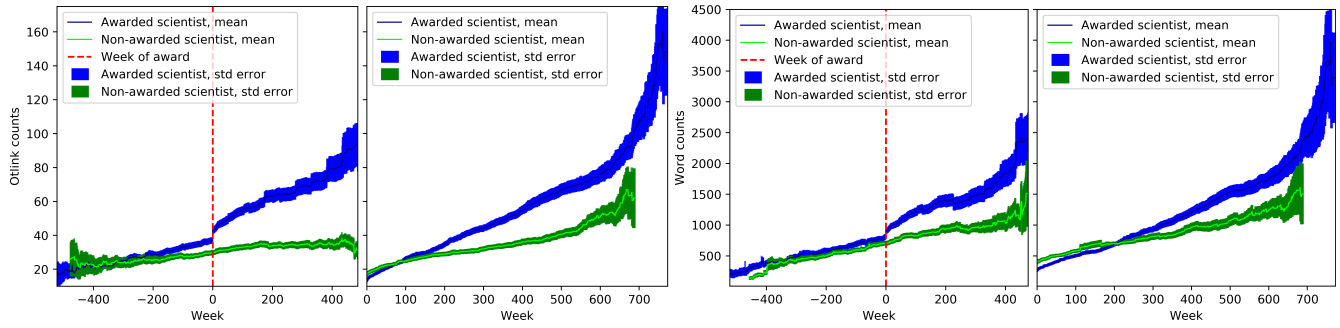
In this work, we compare the temporal patterns of information supply (article creations) and information demand (article views) on Wikipedia for two groups of scientists: scientists who received one of the most prestigious awards in their field and influential scientists that work in the same field but did not receive an award.

Our research highlights that awards function as external shocks which increase information supply and demand about scientists, but hardly affect the demand and supply for information about research topics. Though 95% of articles about scientists have been created before they received an award, information supply is impacted by awards since we find a discontinuity in the growth patterns during the time of the award. After the award, articles about award-winners start to grow much faster than those of non-awarded scientists, while the growth pattern is identical for both groups before that day.

Further, we find interesting differences in the temporal ordering of information supply about scientists and their research topics within the two groups: for award winners information supply about scientists and their research topics is temporally more clustered than for non-award winners. That means for award-winners articles about their research topics are created around the same time as the article about the scientist, while for non-award winners larger time-lags are observed. Award-winners also have a higher probability that interest in them precedes interest in their work, while for non-awarded scientists, 90% (for award-winners only 73%) of the articles about their research topics have been created before the article about the scientist was created. It is not surprising that for both groups the majority of topics were described on Wikipedia before the articles about the scientist was created, since "normal science" is cumulative [9]. That means most scientists work on research topics that have attracted attention in the past. But for award-winners we find more exceptions; 27% of their research

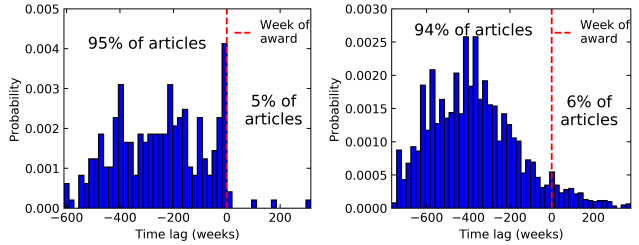
WebSci '18, May 27–30, 2018, Amsterdam, Netherlands

ACM ISBN 978-1-4503-5563-6/18/05.
https://doi.org/10.1145/3201064.3201097



(a) Outgoing link counts (b) Word counts

Figure 1: Cumulative growth of articles on Wikipedia: article length is measured via outgoing links in subfigure (1a) and via word counts in subfigure (1b). The zero point refers either to the week when the scientist was awarded (dashed red line) or the week when the article about the scientist was created (in plots without dashed red line). For the non-awarded scientists we picked a random week out of the range during which awards happened (i.e. between 2008-03-27 and 2015-10-12) as placebo points. One can see a discontinuity in the growth of outlinks and words that is related with the award.



(a) Scientists (b) Research topics

Figure 2: Time lag between the award and the creation date for articles about awarded scientist in subfigure (2a) and for their research topics in subfigure (2b). The zero point on the x-axis refers to the week of the award. Most articles about awarded scientists and their research topics have been created before they received the award.

topics become of interest to the general public after the scientists attracted attention on Wikipedia. One potential explanation is that award winners may innovate new topics or work on relatively new topics. Examples of Wikipedia articles about research topics that were created after the article of the scientists, provide anecdotal evidence for this explanation: Bayesian Networks after Judea Pearl, Public key cryptography after Whitfield Diffie and Martin Hellman, and Tablet PC after Charles P. Thacker.

To our best knowledge, this is the first study that investigates the impact of scientific awards on the production and consumption of information on Wikipedia. We hope that this work is relevant for the Altmetrics community since it sheds light on the temporal dynamics of supply and demand for information about scientists and their research topics online.

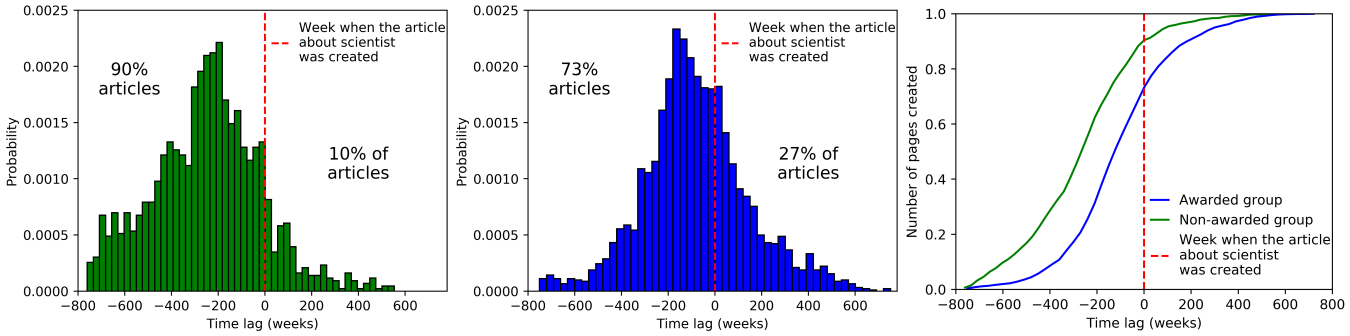
2 DATA AND METHODS

We use Wikipedia article creation dates, edits and views as a proxy for online attention. All data were collected in August 2016. Our code, stopwords list, and datasets are available online[1]. Both datasets of awarded and non-awarded scientist contain the same number of academics from different fields: 57 Physicists, 18 Mathematicians, 18 Computer Scientists, 50 Chemists, 58 Medicine and Physiology researchers, 9 Biologists, and 54 Economists. All together, there are 262 unique researchers in each dataset.

Awarded scientists: This dataset focuses on scientists from the aforementioned fields whose work was honoured through some of the most prestigious academic prizes. We consider the awards between 2008 and 2015, since the Wikipedia page view statistics are not available for earlier years. Within this time frame, we compile a list of winners of the following prizes and awards: Nobel Prize (77 winners), Abel Prize (10), Fields Medal (8), Turing Award (10), IEEE Medal of Honor (8), and International Prize for Biology (9). We also include 163 Thomson Reuters Citation Laureates[2] (23 of whom also received the Nobel Prize), which are selected for outstanding contributions based on the citation impact of their published research. We manually map these winners to the corresponding Wikipedia articles in the English edition, and record their scientific field, gender, award year, and the date when the Wikipedia article was created. The final sample consists of 262 awarded scientists and is available online[3].

Non-awarded Scientists: For a fair comparison, we select a sample of influential, highly cited scientists who worked at the same time, in the same scientific fields as the award winners, using the Thomson Reuters database of Highly Cited Scientists[1]. We use all records between 2001 and 2015 and remove scientists who have

[1]https://github.com/tsennikova/scientists-analysis (accessed Apr. 11, 2018) and https://github.com/gesiscss/scientists-analysis-wikipedia (accessed Apr. 11, 2018)
[2]http://stateofinnovation.thomsonreuters.com/hall-of-citation-laureates (accessed Jul. 28, 2016)
[3]https://github.com/tsennikova/scientists-analysis/blob/master/data/seed/seed_creation_date.json (accessed Apr. 11, 2018)



(a) Non-awarded group (b) Awarded group (c) CDFs of time-lags

Figure 3: Time-lags between creation dates of Wikipedia articles about scientists and their research topics. The zero point on the x-axis corresponds to the week when the article about the scientist was created. Subfigure (3a) shows that the time-lag is negative for 90% of all articles about research topics of non-awarded scientists. This suggests that most articles about research topics have been created before the articles about the non-awarded scientists. For awarded-scientists we see a similar pattern in (3b), but the fraction of articles about research topics that are created after the article about the scientist was created is higher (25%) for awarded scientists than for non-awarded scientists (10%). Subfigure (3c) shows that articles about research topics of non-awarded scientists are created earlier than those of awarded scientists relative to the creation date of articles about scientists.

received an award. Finally, we draw a random stratified sample of 262 academics with the same distribution across scientific fields as in the awarded dataset. We also map these researchers to articles in the English Wikipedia and add information about their scientific field and the date when the Wikipedia article was created (available online[4]).

Scientific Topics: For all researcher in our sample we analyze their Wikipedia article and construct a list of scientific topics related to the scientist. For that, we extract all outgoing links from the articles about scientists in the English Wikipedia. Each of these articles has a category section (found at the bottom of the page) which displays a subject area of the article, and helps readers to navigate through related concepts. We use this concept list and a manually created set of stop words (available online[5]) to remove articles that are related with a scientists but are not related to research areas (e.g. locations, institutions). We evaluate our filtering approach by comparing the algorithmic assessment with a manual assessment for 10 randomly selected articles about scientists and all outgoing links from these articles. The evaluation results show that our filtering method is very effective: the overall accuracy is 0.96, precision is 0.93, and recall is 0.9. Overall, we construct a list of 1,911 topics[6] that are related to awarded scientists and 1,070 topics[7] that are related to non-awarded scientists.

Wikipedia page views: We collect daily page views of all articles about scientists and their research topics. We use page view statistics from the project Wiki Trends[2], which itself is based on the Wikimedia data dumps[3]. Wiki Trends data provides aggregated number of daily visits to Wikipedia articles and all redirects to them, collected from the English edition. In order to eliminate influences of daily and seasonal fluctuations of article views, we normalize the data as follows:

$$\bar{V}_{i,d} = \frac{V_{i,d} * max(M)}{M_d}. \quad (1)$$

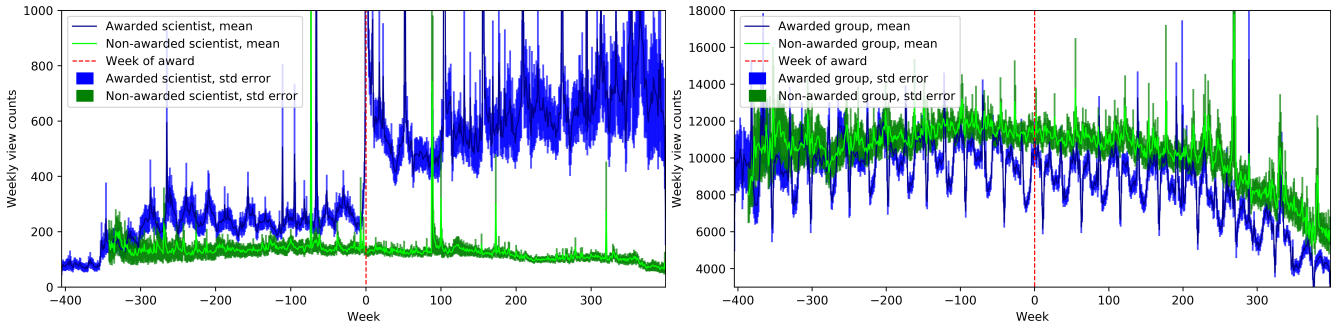
where $V_{i,d}$ refers to the number of visits to an article i on day d, M_d is the number of Wikipedia Main Page views for the same day, and $max(M)$ is the maximum number of Wikpedia Main Page views. We collect page views that happened between 01.01.2008 and 01.05.2016.

3 RESULTS

Information supply: First we explore how collective attention on Wikipedia is affected by external events such as awards, looking at the temporal order of article creations and edits on Wikipedia. Figure 1 shows that articles for awarded and non-awarded scientists grow similarly before the award. But the award creates a discontinuity since articles of awarded scientists start to grow faster than those of their non-awarded colleagues; more hyperlinks and more words are added. This suggests that the award triggers additional information supply, though most articles about awarded scientists and their research topics are created before they receive the award (see Figure 2).

But what came first: the interest in the scientist or the interest in her research? To address this question, we examine the time lag between the article creation about scientists and the scientific topics associated with them. We compare the differences in weeks between the creation dates. The time lag is positive if the topic

[4]https://github.com/tsennikova/scientists-analysis/blob/master/data/baseline/baseline_creation_date.json (accessed Apr. 11, 2018)
[5]https://github.com/tsennikova/scientists-analysis/blob/master/data/neighbors/stop_list.txt (accessed Apr. 11, 2018)
[6]https://github.com/tsennikova/scientists-analysis/blob/master/data/neighbors/seed_neighbors_list_clean_en.json (accessed Apr. 11, 2018)
[7]https://github.com/tsennikova/scientists-analysis/blob/master/data/neighbors/baseline_neighbors_list_clean_en.json (accessed Apr. 11, 2018)



(a) View counts for scientists

(b) View counts for research topics

Figure 4: Weekly view counts for articles about scientists and research topics. The zero point refers to the week when the scientists was awarded (dashed red line). For the non-awarded scientists we picked a random week out of the range during which awards happened (i.e. between 2008-03-27 and 2015-10-12) as placebo points. One can see that the information demand on scientists is clearly impacted by the award, however the interest in research topics associated with the scientists seems to be unaffected.

article was created after the article about the scientist, and it is negative otherwise. Figure 3 shows the probability density function of time lags for both groups of scientists. The zero point on the x-axis refers to the week when the article about the scientist was created. One can see that for both groups most articles about research topics that are associated with a scientist are created before the article about the scientist is created. That means, information supply for research topics usually precedes information supply for scientists. This is not surprising since "normal science" is cumulative [9] and most scientists work on research topics that have attracted attention in the past. But award-winners have a higher probability that interest in them precedes interest in their work. For award winners 27% of articles about research topics related to the scientist have been created after the article about the scientist was created, while for non-award winners only 10% of the articles have been created after the article about the scientist was created. One potential explanation for this difference is that award winners are more likely to innovate new topics or work on relatively new topics and therefore articles about these topics have not yet been created.

Figure 3 also shows that the dispersion of the time-lag distribution for award winners is lower which means that the temporal distances between the creation dates of articles about award-winners and their research topics vary less than for non-award winners. This suggests that interest in scientists and their research topics is more interrelated for award-winners than for non-award winners.

Information demand: So far we have seen that awards impact the production of new information on Wikipedia. Articles about scientists grow faster after they receive an award. However, it remains unclear how the consumption of information is affected by the award. Is the demand for information about scientists and their research topics increasing after they win an award? And how long-lasting is this effect?

Figure 4 shows that the information demand for scientists is impacted by the award, since we see a clear discontinuity in the view counts for articles about scientists who won an award. For non-awarded scientists we pick a random day out of the range during which awards happened (i.e. between 2008-03-27 and 2015-10-12) as placebo points to compute a baseline. The baseline indicates how much change we would expect to see by chance. The discontinuity which we see in Figure 4 clearly goes beyond what we would expect by chance. Also the increased information demand seems to remain rather stable over time. Even 300 weeks after the award, we see that the information demand for awarded scientists is on average higher than those for non-awarded scientists. Interestingly, we see that the information demand for articles about research topics associated with the award-winners seems to be unaffected by the award (cf. Figure 4b).

4 RELATED WORK

Quantifying and predicting scientific success is a topic of high interest for the academic community [4, 6, 11, 14]. While it is clear that online attention to scientists does not always coincide with their academic rigor [12], more research is needed to understand the reasons of such inconsistencies, and the meaning behind them.

Work on collective attention has mainly focused on the consumption of information [8, 15–17] and has shown information consumption correlates with real-world events, such as the spread of influenza [7], box office returns [10] and scientific performance [13].

Only recently researchers started exploring the interplay between information production and consumption. In [5] the authors show that the production of new information on Wikipedia is associated with significant shifts of collective attention measured via article views. That means, in many cases, demand for information precedes its supply. However, unexpected events may lead to almost instantaneously article creations which are followed by a short period of high information demand. A scientific award can be an expected or an unexpected event. Therefore, it is unclear if new articles about scientists will be created on Wikipedia directly after the award, even if no changes in information demand are observed

before the award. Our work shows that in most cases articles about scientist and research topics precede the award. However, we see that awards boost the demand for information about scientists and that the increased demand lasts over the next few years.

5 DISCUSSION

How does an award impact information supply and consumption online? If awards would be totally unexpected and hit scientists randomly, they would lead to almost instantaneous article creations which would be followed by a short increase in information demand [5]. Our work shows a different pattern and suggests that awards are probably not so unexpected and have long term effects. 95% of the articles about scientists are created before they receive an award, but information supply is impacted by awards since articles about award-winners grow faster than those of non-awarded scientists. Also information supply is impacted by the award since we find a discontinuity in the view counts in the week when the prize was awarded. Interestingly the increased information demand is rather stable over time. Even five years after the award, we see that the information demand for awarded scientists is on average higher than the demand for non-awarded scientists.

The discontinuity which we see during the week when the prizes are awarded suggests that awards may have a causal effect on information production and consumption. However, to establish a hard causal link future research is necessary since other factors that correlate with awards may exist and confound our analysis.

We also find interesting differences between the two groups of scientists when looking at the information production side. For award-winners articles about them and their research topics are temporally more clustered than for non-award winners. Award-winners also have a higher probability that interest in them precedes interest in their work. One potential explanation is that award winners are more connected with their research topic since they may innovate new topics or work on relatively new topics. Examples of topic articles that were created after the article of the scientist provide anecdotal evidence for this explanation: Bayesian Networks after Judea Pearl, Public key cryptography after Whitfield Diffie and Martin Hellman, and Tablet PC after Charles P. Thacker. However, further research is necessary to explore the different types of relationships between scientists and their research topics that will lead to the creation of a hyperlink on Wikipedia.

6 CONCLUSIONS

The goal of this work was to understand the impact of awards on the production and consumption of information about scientists and their research topics on Wikipedia.

Our work shows that (i) scientists who win a prestigious prize attract more attention afterwards (i.e., information supply and demand increases but only for articles about scientists); (ii) information supply for award winners and their research topics is temporally more clustered than for non-award winners; and (iii) information supply for award winners is more likely to precede information supply for their research topics compared to non-award winner.

For future work it would be interesting to extend this group level analysis with an individual level analysis and further investigate the different types of relationships between scientists and their research topics that may lead to a hyperlink on Wikipedia. We also collected gender information about scientists and plan to compare the information supply and demand for female and male award-winners and influential scientists that did not receive an award in future work.

7 ACKNOWLEDGMENTS

This work is part of the DFG-funded research project **metrics* (project number: 314727790). Further information on the project can be found at https://metrics-project.net.

REFERENCES

[1] 2016. Thomson Reuters database of Highly Cited Scientists. http://hcr.stateofinnovation.thomsonreuters.com/page/archives. Accessed: 2016-06-28.
[2] 2016. Wiki Trends Project. http://www.wikipediatrends.com/. Accessed: 2016-06-28.
[3] 2016. Wikimedia Data Dumps. https://dumps.wikimedia.org/. Accessed: 2016-06-28.
[4] Judit Bar-Ilan, Stefanie Haustein, Isabella Peters, Jason Priem, Hadas Shema, and Jens Terliesner. 2012. Beyond citations: Scholars' visibility on the social Web. (2012). http://arxiv.org/abs/1205.5611 17th International Conference on Science and Technology Indicators, Montreal, Canada, 5-8 Sept. 2012.
[5] Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2015. The production of information in the attention economy. *Scientific Reports* 5 (2015), 9452. https://doi.org/10.1038/srep09452
[6] Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. 2018. Science of science. *Science* 359, 6379 (March 2018), eaao0185. https://doi.org/10.1126/science.aao0185
[7] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232 (Feb. 2009), 1012–1014. https://doi.org/10.1038/nature07634
[8] Nathan Oken Hodas and Kristina Lerman. 2012. How Visibility and Divided Attention Constrain Social Contagion. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust (SOCIALCOM-PASSAT '12)*. IEEE Computer Society, Washington, DC, USA, 249–257. https://doi.org/10.1109/SocialCom-PASSAT.2012.129
[9] Thomas S. Kuhn. 2012. *The Structure of Scientific Revolutions: 50th Anniversary Edition*. University of Chicago Press, Chicago 60637.
[10] Márton Mestyán, Taha Yasseri, and János Kertész. 2013. Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. *PLOS ONE* 8, 8 (Aug. 2013), e71226. https://doi.org/10.1371/journal.pone.0071226
[11] Orion Penner, Raj K. Pan, Alexander M. Petersen, Kimmo Kaski, and Santo Fortunato. 2013. On the Predictability of Future Impact in Science. *Scientific Reports* 3 (Oct. 2013), 3052. https://doi.org/10.1038/srep03052
[12] Anna Samoilenko and Taha Yasseri. 2014. The distorted mirror of Wikipedia: a quantitative analysis of Wikipedia coverage of academics. *EPJ Data Science* 3, 1 (Dec. 2014), 1. https://doi.org/10.1140/epjds20
[13] Hua-Wei Shen and Albert-László Barabási. 2014. Collective credit allocation in science. *Proceedings of the National Academy of Sciences* 111, 34 (2014), 12325–12330. https://doi.org/10.1073/pnas.1401992111
[14] Roberta Sinatra, Dashun Wang, Pierre Deville, Chaoming Song, and Albert-László Barabási. 2016. Quantifying the evolution of individual scientific impact. *Science* 354, 6312 (2016). https://doi.org/10.1126/science.aaf5239
[15] Mike Thelwall, Kayvan Kousha, Katrin Weller, and Cornelius Puschmann. 2012. *Chapter 9 Assessing the Impact of Online Academic Videos*. Emerald Group Publishing Limited, Chapter 9, 195–213. https://doi.org/10.1108/S1876-0562(2012)0000005011
[16] Lilian Weng, Alessandro Flammini, Alessandro Vespignani, and Filippo Menczer. 2012. Competition among memes in a world with limited attention. *Scientific Reports* 2 (March 2012), 335. https://doi.org/10.1038/srep00335
[17] Fang Wu and Bernardo A. Huberman. 2007. Novelty and collective attention. *Proceedings of the National Academy of Sciences* 104, 45 (Nov. 2007), 17599–17601. https://doi.org/10.1073/pnas.0704916104