Natural Language Engineering **23** (5): 807–812.

doi:10.1017/S1351324917000286

Emerging trends: Inflation

KENNETH CHURCH
IBM
e-mail: kenneth.ward.church@gmail.com

(*Received 14 June 2017; revised 25 June 2017*)

Abstract

Our field has enjoyed amazing growth over the years. Is this a good thing or a bad thing, or just a thing? Good: Growth sounds good. It is hard to imagine a politician arguing against jobs. There are more people working in the field than ever before, and they are publishing more and more, and creating more and more value. What could be wrong with that? Bad: Whatever you measure you get. We are all under too much pressure to publish too much too quickly. Students are graduating these days with more publications than what used to be expected for tenure. So many people are publishing so much that no one has time to think great thoughts, or take time to learn about things that may not be directly relevant to the next publication. Neutral: Inflation is a fact of life. There are long-term macro trends on publication rates that are beyond our control. These trends hold over tens and hundreds of years, and will continue over the foreseeable future.

1 Is inflation good, bad or neutral?

Growth sounds good. It is hard to imagine a politician arguing against jobs. There are more people working in the field than ever before, and they are publishing more and more, and creating more and more value. Fortune 500 companies (e.g., Google, Apple, Amazon, Facebook, IBM) are paying attention to what we do. What could be wrong with that?

I worry that we are all under too much pressure to publish too much too quickly. (Whatever you measure you get.) Students are graduating these days with more publications than what used to be expected for tenure. So many people are publishing so much that no one has time to think great thoughts, or take time to learn about things that may not be directly relevant to the next publication. I recently attended a workshop where senior people were encouraging people to take more time to write better papers, and students were pushing back, explaining the current realities: publish (every month) or perish.

Fernando Pereira recently published a blog post[1] on the history of the field in response to a post by Yoav Goldberg.[2] I will bring the discussion back to growth

[1] http://www.earningmyturns.org/2017/06/a-computational-linguistic-farce-in.html

[2] See https://medium.com/@yoav.goldberg/an-adversarial-review-of-adversarial-generation-of-natural-language-409ac3378bd7 and links therein.

rates shortly, but let me first mention a few highlights that resonate with some of my Emerging Trends pieces.

Yoav Goldberg started the discussion by complaining about a particular arXiv paper, and more broadly about a tendency for certain 'deep learning' papers to overstate results in areas that they do not know much about. I will not talk about the particular paper in question, but we have all been frustrated by papers that claim to have more than they have, and disrespect a field with a lack of knowledge/appreciation for the background literature. I agree that it is ok to reject the classics, but not without reading them.

That said, as I mentioned in my first Emerging Trends piece on the next generation (Church 2016), the future of our field depends on the next generation. Kuhn observes that young people are often the early adopters, the first to see what is about to happen, and those with the most to lose (the establishment) tend to be the most resistant to change. We are counting on the next generation to rock the boat, and the last generation to keep it from flipping over. I agree with Yann LeCun[3] when he encourages the next generation to do what they need to do to make sure that the field continues to prosper well after we are gone.

LeCun also defends arXiv for reasons that are similar to my most recent Emerging Trends piece (Church 2017) where I pointed out that arXiv ranks highly on Google Scholar's leaderboard, well above all the journals in our field including this journal. The leaderboard is based on h-index, which counts quality (papers with lots of citations), with no penalty for quantity (unlike impact factor). I agree with LeCun that we should focus on the best papers in a venue (as h-index does), and do our best to ignore the rest. Although I share much of Goldberg's frustration, I know deep down that it is rather pointless to get too worked up about things that are beyond my control. No matter what we do, there will always be a few good papers that are well worth reading, and many more that are not.

In addition to those more controversial points, I would like to focus on a less controversial paragraph in Pereira's post that ends with a statement about growth rates:

> The empiricist ascendancy was fortunate (or prescient) in riding the growth of computing resources and text data that also enabled the Web explosion and the flooding of all this work with new resources for funding research, software development, and corpus creation. The metrics religion helped funders sell progress to the holders of the purse strings, and there were real (if not as extensive as sometimes claimed) practical benefits, especially in speech recognition and machine translation. As a result, the research community grew a lot (my top-of-the-head estimate is around 5x from 1990 to 2010).

I am particularly interested in the '5x.' Pereira uses growth rates in a number of places to argue that certain changes in direction have been good for the field.

The fact that there was a revival of empiricism in the 1990s is fairly uncontroversial. Most would agree, moreover, that the revival of empiricism had a big impact on the field, though there was disagreement, especially at first, about whether the impact was good or bad. I have published similar comments myself (Church 1993;

[3] https://www.facebook.com/yann.lecun/posts/10154498539442143

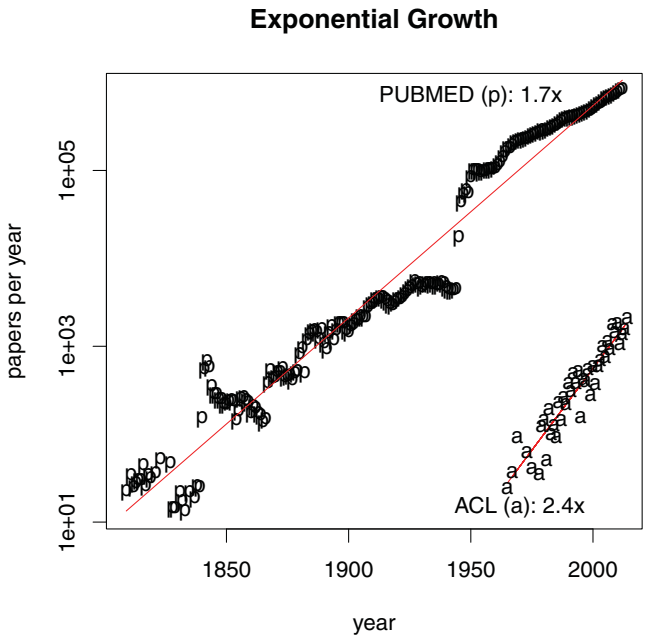

Fig. 1. (Colour online) Scatter plot of publications over time shows exponential growth. The plot shows two datasets: PUBMED (labeled 'p') and ACL Anthology (labeled 'a'). Regression lines are shown in red. Publications are increasing at a growth rate of 1.7 times per decade for PUBMED and 2.4 times per decade for ACL.

Church 2011), but I have never attempted to quantify the magnitude of the impact with a number like '5x.'

To make sense of a number like '5x,' what is a reasonable baseline to compare that to? Pereira is assuming that 5x is a big deal, but how much growth should we expect to see over 20 years (all other things being equal)? Is 5x more than typical growth?

It is well known that publication rates have been growing exponentially for decades; see Bornmann (2015) and references therein. Figure 1 shows this trend for two datasets: ACL Anthology (labeled 'a') and PUBMED (labeled 'p'). The growth rates are 2.4 per decade for ACL Anthology and 1.7 for PUBMED. That corresponds to a doubling rate of 8.1 years for ACL Anthology and 12.5 years for PUBMED. Over 20 years, we would expect to see 5.6 more ACL papers and 3.0 more PUBMED papers. Thus, Pereira's 5x does not appear to be significantly more than we would expect to see over a random twenty-year period, at least for the ACL Anthology.

The PUBMED data span 100s of years, an order of magnitude more than the ACL Anthology. The additional time makes it easier to see differences between more productive decades and less productive decades. There was a slow down during the great depression in the 1930s, followed by a large boom starting in 1945 immediately after the war. Smaller slow downs can also be observed for other wars such as the American Civil War and WW I. Macro statistics such as these show dramatic residuals from the trend for significant events such as major recessions and wars,

https://doi.org/10.1017/S1351324917000286 Published online by Cambridge University Press

but it is harder to make strong statements about relatively less important events such as the 1990s revival of empiricism.

It has been argued in Bornmann (2015) that growth rates are accelerating, though my attempts to reproduce their results with PUBMED and ACL anthologies were unsuccessful. They report a large increase in growth rates of the Web of Science from essentially flat in 1650 (1.0x per decade) to 2.2x per decade at present. Expressed in terms of doubling rates, they estimate that it currently takes less than nine years to double the size of the literature, considerably less than 231 years that it took to double the size of the literature back in 1650. Their doubling rate of nine years is roughly comparable to the doubling rates reported above of 8.1 years for ACL and twelve years for PUBMED. On the other hand, I see little evidence in my data sets for a different doubling rate at the beginning than at the end.

With or without acceleration, the inflation rates in Figure 1 are substantial. What could be causing so much inflation? The literature appears to be growing faster than the value of money. $1 in 1990 is worth between $1.67 and $2.50 in 2010,[4] whereas we reported above, for the same time period, an increase of 5.6x for ACL papers and 3.0x for PUBMED. Perhaps, the worldwide R&D budgets have increased in terms of inflation adjusted investments, but I suspect that there are additional factors behind the inflation rates in Figure 1.

I am particularly concerned about the number of publications per person per year. As mentioned above, I worry that we are all under too much pressure to publish too much too quickly. Publish or perish is nothing new. There have always been promotion processes and annual review processes that evaluate individual researchers, as well as institutions, based on a number of factors including publications. The UK, for example, has its Research Excellence Framework,[5] and there are similar processes elsewhere.[6] I do not know enough about these processes to criticize them, but it is possible that such mechanisms could be inherently inflationary.

Such review processes may lead to unintended consequences. When I was at Hopkins, which is largely a medical school, publications did not count unless they were indexed by the Web of Science. The Web of Science may be ok for some fields like Medicine, but less so for our field. The Web of Science excludes about half of my publications (and most of my citations). The Web of Science prefers publications in obscure journals behind pay walls over venues that people read and respect (and cite).

Another unintended consequence are vanity presses. My inbox these days seems to be full of spam email advertising more and more opportunities to publish in more

[4] https://www.measuringworth.com/uscompare/result.php?year_source=1990&amount=1&year_result=2010

[5] https://en.wikipedia.org/wiki/Research_Excellence_Framework

[6]
- https://en.wikipedia.org/wiki/Excellence_in_Research_for_Australia
- http://www.ugc.edu.hk/eng/ugc/activity/research/rae/rae2014.html
- https://www.knaw.nl/en/news/publications/standard-evaluation-protocol-2015-2013-2021?set_language=en

and more venues that I have never heard of. There is an entertaining story about how one author attempted to address such spam.[7] It is easy to blame vanity presses for doing what they do, but they would not exist if people did not feel so much pressure to publish. It cannot be good for the field if many of us feel the need to publish (every month) or perish.

I really respect review processes that encourage people and institutions to publish less. Bell Labs actually believed that less is more. Doug McIlroy, the boss of the people who invented Unix, advocated the less-is-more Unix Philosophy.[8] McIlroy had no time for clutter. I heard him talk about his favorite programmer that deleted more code than he wrote:[9] 'The real hero of programming is the one who writes negative code.' There was so much talk about 'less is more' that it led to some entertaining puns.[10]

Like most institutions, Bell Labs had an annual review process. But this process valued quality over quantity (and accomplishments over activities). My most memorable review was after a particularly productive year. My boss suggested that a paper a month was too much. He said that they stop counting after four papers. They figure that if you publish more than that, they cannot be that good. I could have tried to push back with an argument about how they really were that good, but it would have been pointless to go there. They felt the need to make a statement about quality (and not about quantity). It would be too easy for an institution (or a research community) to consume all available resources creating clutter, and they wanted to make sure that didnot happen by discouraging quantity (even if my papers really were that good).

I do not know how we can address the clutter in our field. I do not think it is productive to complain about a particular bad paper in arXiv. There are too many bad papers and too much clutter to remove each instance, one at a time.

A more effective approach is to address incentives. Why do people feel the need to publish so much? Bad review processes are like bad papers. There will always be a few good papers that are well worth reading, and many more that are not. So too, there will always be a few good review processes (like Bell Labs) that are well worth paying attention to, and many more that are not. It is more productive to praise the best than to complain about the rest.

References

Church, K. 2016. Emerging trends: the next generation. In *Natural Language Engineering*, vol. 26, no. 6, pp. 997–980. Cambridge University Press.

Church, K. 2017. Emerging trends: I did it, I did it, I did it, but…. In *Natural Language Engineering*, vol. 23, no. 3, pp. 473–480. Cambridge University Press.

Church, K. 2011. A pendulum swung too far. *Linguistic Issues in Language Technology* **6**(5).

[7] https://www.theguardian.com/australia-news/2014/nov/25/journal-accepts-paper-requesting-removal-from-mailing-list

[8] https://en.wikipedia.org/wiki/Unix_philosophy

[9] https://en.wikipedia.org/wiki/Douglas_McIlroy

[10] https://en.wikipedia.org/wiki/Less_(Unix)

https://doi.org/10.1017/S1351324917000286 Published online by Cambridge University Press

Church, K. and Mercer, R. 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics* **19**(1), 1–24.
Bornmann, L. and Mutz, R. 2015. Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology* **66**(11), 2215–2222.