Response to Comment on “Quantifying long-term scientific impact”

Dashun Wang,1 Chaoming Song,2 Hua-Wei Shen,3,4 Albert-László Barabási2,5

Wang, Mei, and Hicks claim that they observed large mean prediction errors when using our model. We find that their claims are a simple consequence of overfitting, which can be avoided by standard regularization methods. Here, we show that our model provides an effective means to identify papers that may be subject to overfitting, and the model, with or without prior treatment, outperforms the proposed naïve approach.

Response 1: Papers with large parameter sets are a simple consequence of overfitting during the likelihood estimation. To prevent overfitting, one should follow standard procedures (3) by applying regularization methods. To illustrate this practice, we applied a conjugate prior, finding that overfitting was avoided (4). We repeated our analysis on the same corpus as described by Wang, Mei, and Hicks (5) (papers published in 1980 in the Physical Review data set that received at least 10 citations within the first 5 years, resulting in 681 papers). We obtained parameters for these papers using 10 years as training data ($a = 4.7137$ and $b = 5.6273$) and plotted the correlation between parameter $\lambda$ and $\mu$ (Fig. 1A), finding that the overfitting issue is completely avoided. We also compared the model with two additional baseline models that we could not include in (2)—both are more competitive than the naïve approach suggested in (1)—finding that our model consistently beats competing methods under different performance metrics, including the one used in (1) [mean absolute percentage error (MAPE)].

Response 2: This is the result of a few papers with large errors, caused by the overfitting mentioned in response 1. The WSB model offers a way to detect papers that may be subject to overfitting. Upon filtering these papers, the large errors reported in (1) disappear and the WSB model outperforms the naïve approach.

Response 3: The performance of the WSB model with prior is still worse than the naïve approach. Wang, Mei, and Hicks's experiments were based on an incorrect set of prior parameters. The WSB model with prior consistently outperforms the naïve approach.

Response 4: We repeated the analysis on the same corpus as the one used by Wang, Mei, and Hicks (the same as in Response 1). We find that for both 5 and 10 years training periods, the WSB model with prior also yield much lower MAPE (0.204 versus 0.305 for 10 years and 0.396 versus 0.504 for 5 years). We also plotted the quantities shown in the figure 1 of (1), finding substantial differences in both parameter correlations (Fig. 1A) and predicted citations (Fig. 1C). Wang, Mei, and Hicks also used other evaluation metrics, such as Spearman correlation and percentage of correctly identified top 10% papers. We therefore evaluated these two metrics as well, finding that the WSB model with prior consistently outperforms the naïve approach for both metrics using 5 and 10 years of training. We repeated our analysis on other sets of papers as well, such as papers published in the 1970s, finding consistent results. The results in (1) were based on an incorrect set of prior parameters. Indeed, they applied priors learned on review papers in Reviews of Modern Physics to predict research papers published by Physical Review. It is reasonable to assume that the WSB model with prior outperforms the naïve approach for a wide range of prior parameters, as long as they are within a reasonable range, because the naïve approach reduces to a special case of the WSB model with prior with a trivial prior parameter $\beta \rightarrow = (4)$.

Taken together, the large prediction errors reported by Wang, Mei, and Hicks is a consequence of overfitting, which can be avoided using standard regularization methods. They misinterpreted the true message of the paper. In our view, the main message of (2) lies in the uncovered regularity of a complex evolving system that previously was perceived as noisy and unpredictable. Indeed, the proposed model is a minimal citation model that captures all quantifiable mechanisms known to date to affect citation histories. Hence, it represents a fundamental building block on which one needs to add industrial-quality implementation protocols, which could lead to a more robust and accurate citation prediction tool. Therefore, judging our results on the quality of the implementation is like judging the laws of thermodynamics on the performance of the cars a particular company can build.
Fig. 1. (A) Correlation between parameters $\lambda$ and $\mu$ using the WSB model with prior, indicating that applying a prior resolves the issue of overfitting completely. The $\mu$ parameters obtained by Wang, Mei, and Hicks are smaller than ours. We suspect this is due to unit conversion. That is, they might have used years instead of days as a time unit. (Inset) Same correlation by using the WSB model, documenting the existence of outliers also observed in (1) due to overfitting. (B) Comparison between the predicted and the real citations for the WSB model in cases where $s/c_{\text{pred}} < 1$, accounting for 87.2% of the papers. The lack of outliers demonstrates that the citation envelope provides an effective means to filter out cases that are affected by overfitting. (C) Comparison between predicted and real citations for the WSB model with prior. Comparing (B) and (C), we find that applying prior improves the predictive power of the WSB model. (D) Comparison between the predicted and the real citations for the naïve approach, documenting that it systematically underestimates the future citations. We used 10 years of training for results in (A) to (D). All conclusions continue to hold if we use 5 years of training.
Response to Comment on "Quantifying long-term scientific impact"
Dashun Wang, Chaoming Song, Hua-Wei Shen and Albert-László Barabási

Science 345 (6193), 149.
DOI: 10.1126/science.1248961