Cost of Research and Education Activities in US Colleges - Scalability, Complementarity, and Heterogeneous Efficiency

Authors: Hajime Shimao[a,b] (corresponding author) Xiaoxiao Li[c], Michael Holton Price[a], Christopher P. Kempes[a]

Affiliations: a. Santa Fe Institute, b. McGill University, c. Villanova University **Contact:** hajime.fr@gmail.com

Introduction

Universities in the United States are remarkably diverse in their efficiency, both in terms of research output and educational achievement. In the existing literature, the heterogeneity is under-explored due to the lack of appropriate data and methodological limitations. In this paper, we address this by exploiting a newly consolidated dataset and adopting a neural-network-based method to infer cost functions for universities. Our analyses reveal that there are substantial efficiency differences across universities. Particularly, we show that while both research and education outputs generally exhibit an economy of scale, their scalability largely depends on the size and other institutional characteristics. Similarly, research and education activities are complementary to each other (economy of scope) only when the scales of productions are small to medium. Furthermore, the empirical cost isoclines of universities can be non-convex which leads to important policy implications, including diverse optimal portfolios and specialization. In short, our fully data-driven analysis suggests that model assumptions need empirical validation.

Methodology

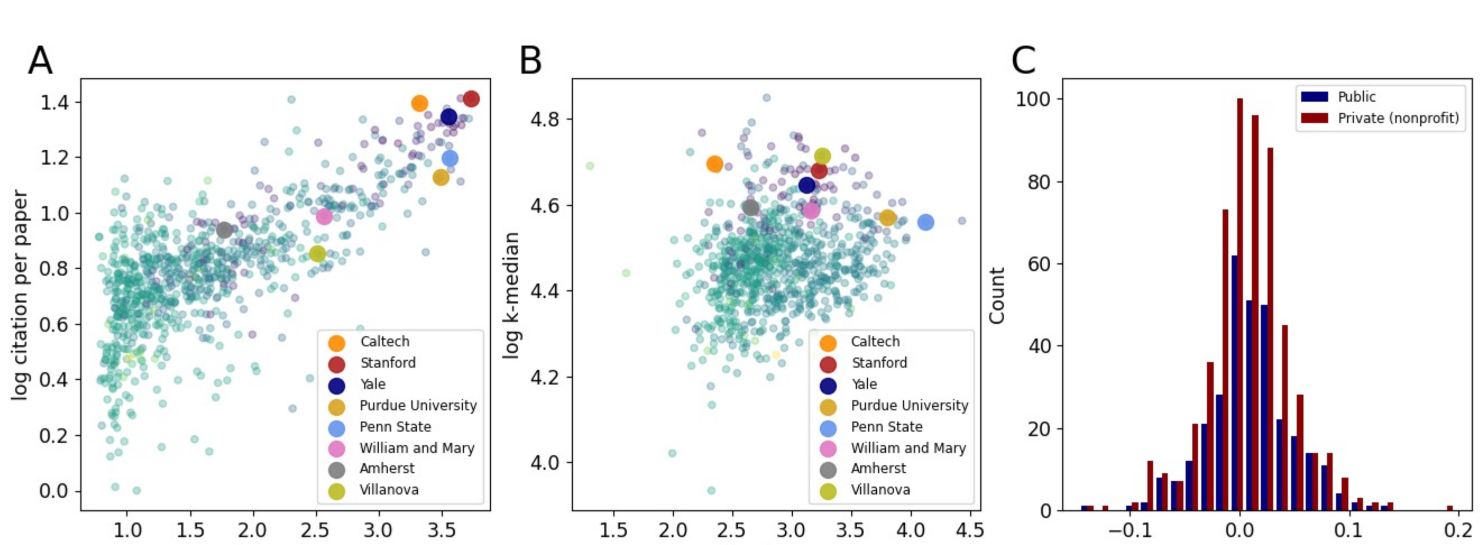
Our framework combines modern deep neural network techniques with recent advances in "interpretable machine learning" algorithms. Specifically, we estimate cost as a nonlinear function encoded by a neural network without pre-specifying parametric assumptions; this allows complete flexibility in the functional form. Further, we utilize Bayesian optimization to find the architecture and hyperparameters of the neural network that properly model the non-linearity of the cost function without overfitting the data. We thus allow for heterogeneity from both observable characteristics and unexplained factors, such that the cost function captures the diversity exhibited by schools and can offer effective policy implications specific to each institution. In short, to model the total cost, we replace a translog function with the following flexible functional form:

$$log_{10}(C_{it}) = f(y^r_{it}, y^e_{it}, \mathbf{z}_{it}) + e_{it}$$

where $f(.)$ represents the function represented by the neural network. This flexible cost function allows analyses of the scalability and complementarity of research and education that are not possible with restrictive specifications of the cost function, such as the translog cost function.

Data – Heterogeneity

We construct comprehensive measures of university outputs for both research and education. **UnivProd dataset:** (1) Microsoft Academic Graph (MAG) which contains detailed information on academic publications, including author and institutional affiliation. (2) Mobility Report Card collected by Chetty et al (2017) which reports income data for graduates and parents by university. (3) IPEDS data which provide budgetary information. We describe the generation of the dataset in a separate paper (Price et al. 2022). Here, we offer the first analysis that uses this dataset.



A,B: The distribution of the quantity-quality in research and educational outputs. The colors are based on the Barron's selectivity measure for small dots (darker color indicates more selective), and large dots indicates observations for a set of arbitrarily selected schools.

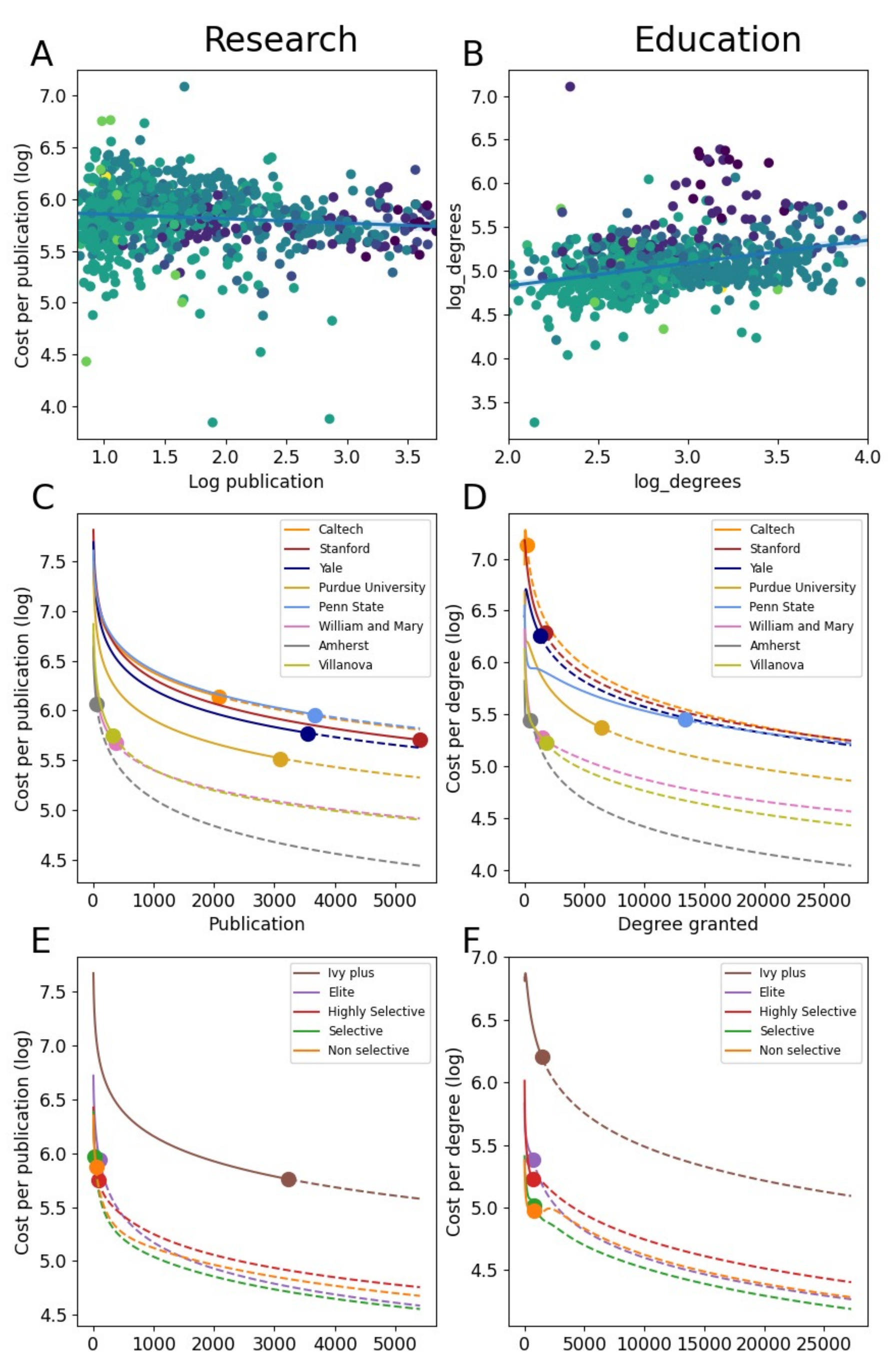
C: The distribution of the unobserved cost shifter μ. Most of the value range between -.1 and .1, indicating that total cost may vary $\pm 10\%$ based on factors not observed in the dataset.

Results – Scalability

We would like to assess how much a school is paying for a unit of publication/degree. To this end, we compute the average incremental cost (AIC) of research and education outputs, respectively, as

$$AIC_i^r(y^r, y^e) = \frac{C_i(y^r, y^e) - C_i(0, y^e)}{y^r}$$

$$AIC_i^e(y^r, y^e) = \frac{C_i(y^r, y^e) - C_i(y^r, 0)}{y^e}$$



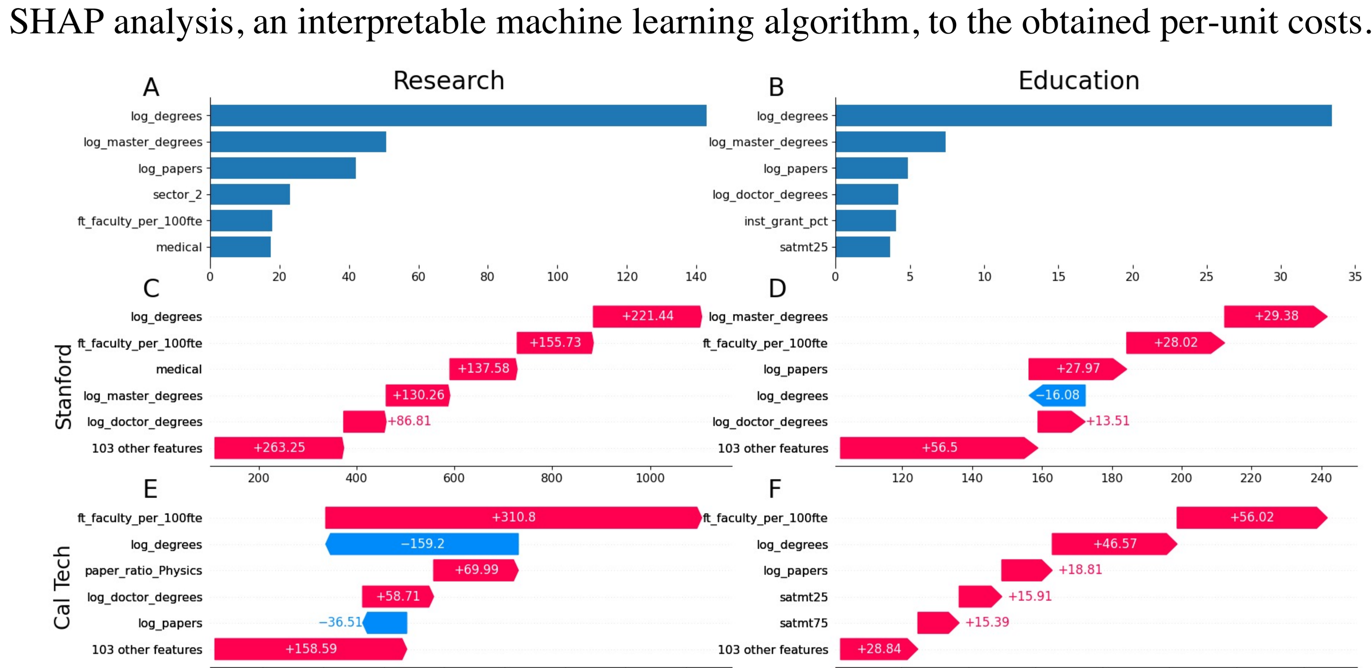
A, B: Scatter plot of the scale of production (x-axis) and the estimated AIC (y-axis) in research (A) and education (B). The correlation is weakly negative (-.04) for research and positive for education (.26).

C, D: Simulated AIC curves of research (C) and education (D) outputs for the example schools. The dots indicate their actual production levels. The lines for Williams and Mary, Amherst, and Villanova are below other more research-intense schools, indicating that their per-unit costs are lower at a given scale. However, the scalability is not fully utilized.

E, F: Simulated AIC curves of research (E) and education (F) outputs when control variables are set to be the median values of each selectivity tier. Lower tier schools pay lower per-unit costs if the scales are the same, but their research does not exploit the scale merit.

Results – SHAP Analysis

To further infer what institutional characteristics can explain the observed cost efficiency, we apply the SHAP analysis, an interpretable machine learning algorithm, to the obtained per-unit costs.

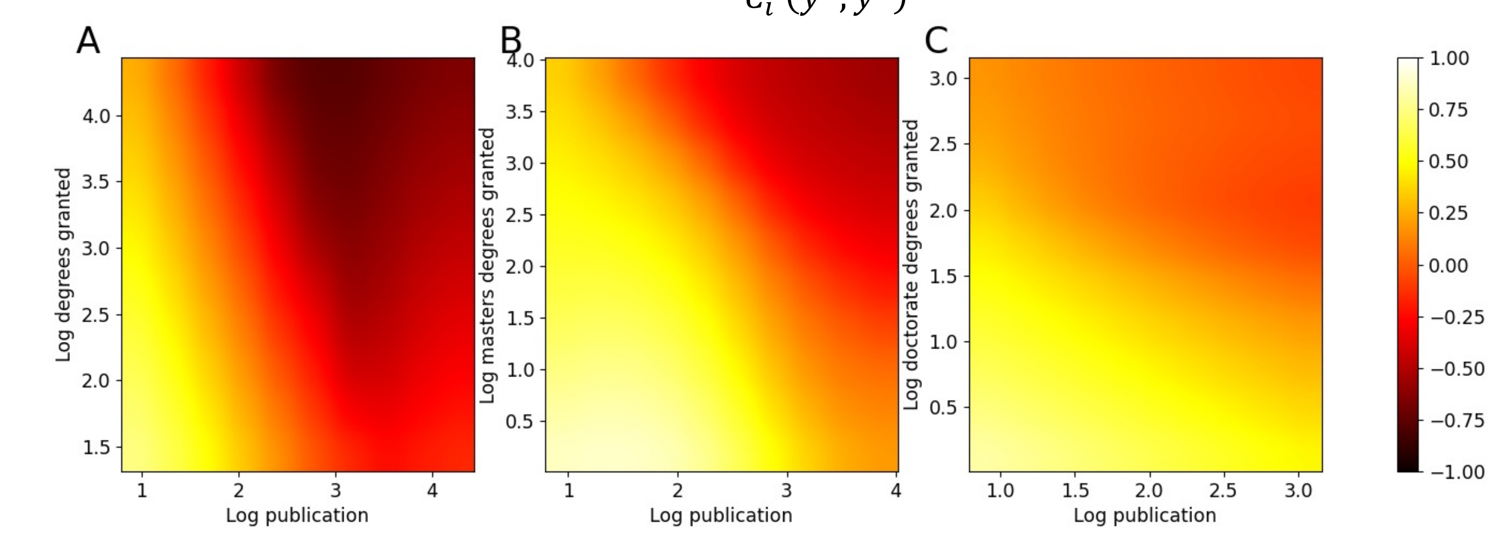


A, B: The mean absolute importance of features on AICs of research (A) and education (B). **C, D:** SHAP analysis on the AICs of Stanford. The size of undergraduate education (log_degrees) is the most influential feature increasing its research cost (C), and the fourth important feature reducing its educational cost (D). **E, F:** SHAP analysis on the AICs of Caltech. Contrary to Stanford, its small-scale education reduces the research cost but increases the education cost.

Results – Complementarity

Is there complementarity between research and education productions? What's the efficiency gain in operating these two activities within the same organization? To explore this, we compute a metric of complementarity (i.e. economy of scope) between research and education productions as

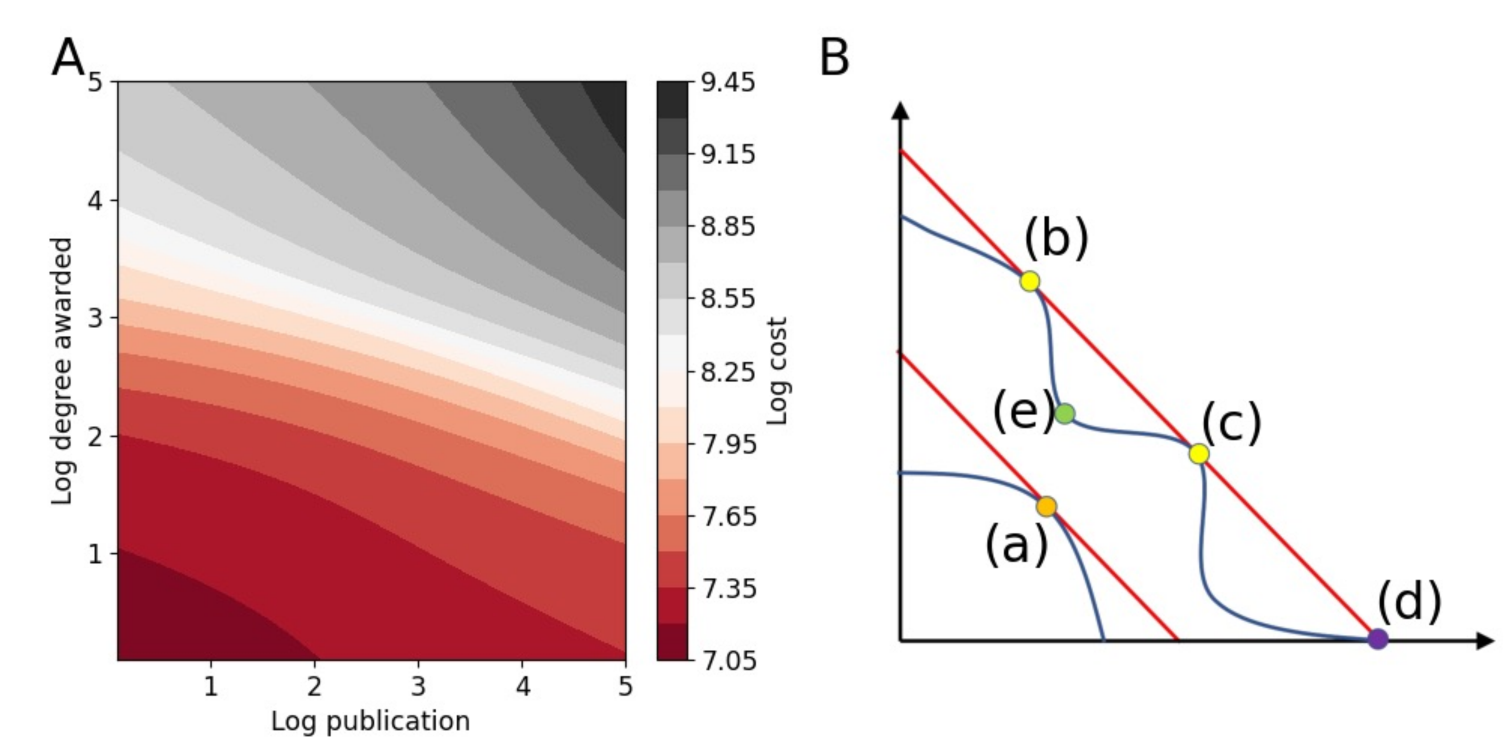
$$SC_i = \frac{C_i(0, y^e) + C_i(y^r, 0) - C_i(y^r, y^e)}{C_i(y^r, y^e)}$$



Heatmap depicting the estimated complementarity between research and education (A undergraduate, B master, C doctoral). x-axis and y-axis indicate the size of the research and education respectively. All the other control variables are set to be the median of the dataset. Brighter color means larger positive complementarity between the two activities.

A: Undergraduate education and research are complementary only when the size is small. When the size of both outputs is large (top-right), it can become negative. **B,C:** Graduate educations are more complementary to research activities.

Results – Theory Implication



A: Empirical cost isoclines along research (x-axis) and education (y-axis) output. As one can see, the cost isoclines are not convex in all regions, which has several important implications. (1) There may exist *multiple optima* (Figure **B** point b and c). (2) There may exist a *corner solution* where a university focuses on a single output (point d). (3) Some points (such as e) are never optimal under any preference weight. All the possibilities challenge the traditional assumption of rational choice and can lead to non-trivial policy implications. Model assumptions need empirical validation.

Conclusion

In this paper, by applying modern machine learning techniques to our novel data, we attempt to minimize the assumptions imposed by researchers and our fully data-driven analysis reveals that questions of scalability and complementary are highly conditioned on features of universities.

Our resulting dataset, **UnivProd**, will be freely available soon. We describe the generation of the dataset in a separate paper (Price et al. 2022). Please contact the corresponding author, Hajime Shimao (hajime.fr@gmail.com), for questions as well as access to the dataset and our manuscripts.