Which universities lead and lag? Toward university rankings based on scholarly output

Daniel Ramage and Christopher D. Manning
Computer Science Department
Stanford University
Stanford, California 94305
{dramage,manning}@cs.stanford.edu

Daniel A. McFarland
School of Education
Stanford University
Stanford, California 94305
dmcfarla@stanford.edu

1 Introduction

Science policy makers, university administrators, funding agencies, and prospective students all rely on many factors when deciding which academic institutions to become involved with. Organizations have stepped in to provide information to support such decision makers. From US News and World Report rankings[1] to the recently released National Research Council report on academic institutions [4], there is great interest in generating objective benchmarks of academic institutions. Traditionally, these benchmarks have focused on the inputs associated with each institution—amount of money raised, SAT scores of incoming students, number of grant dollars and research staff, etc—or on the reputation of those institutions as judged by their peers. Yet by their nature, academic institutions produce a great deal of output, usually in the form of the text of academic publications or dissertations. Such text-rich datasets tend to be overlooked in quantitative analysis of institutional performance because making effective, quantitative use of text is a challenging problem. In this study, we analyze these same institutions from a new perspective: scoring institutions by how much each institution looks like the future of academia, judged quantitatively from the text of each institution's PhD dissertation abstracts.

PhD dissertation abstracts are an excellent basis for the study of academic institutions because they reflect the entire academic output of a university, as seen through its graduating students. By contrast, common databases of academic publications tend to be biased toward particular domains, such as biomedical research for ISI or computer science for CiteSeer. And these databases do a particularly poor job of representing academic output in areas such as the humanities, where books are the primary currency of academic scholarship. By contrast, every graduating PhD student produces a dissertation reflecting multiple year's work in what ultimately results in a new contribution to human knowledge, as judged by the faculty members advising that student.

In this work, we ask a simple question: based on the content of PhD dissertation abstracts, which universities are future-leaning, or leading, and which are past-leaning, or lagging. We consider a university to be future leaning if, in any given year, its dissertations tend to be more similar to academia's future than to its past. Conversely, a university is past-leaning if, in any given year, its dissertations tend to be more similar to dissertations from academia's past. Universities can be compared and ranked based on the difference between a university's similarity to the future and its similarity to the past, insofar as we trust our similarity measures. We believe that this simple framing presents a new way to deeply analyze the institutions of academia from a content-driven, output-oriented perspective that can form the basis of a wide range of tools to explore more fine-grained social science questions, including quantifying the roles played by the various input factors measured in the traditional rankings and how those translate (or don't translate) to academies that look more or less like the future.

[1] http://www.usnews.com/rankings

2 Dataset

Shortly after completion, almost all PhD dissertations in the United States are filed in the UMI database maintained by ProQuest.[2] ProQuest is a private company designated by United States Library of Congress as the collection agency for all PhD dissertations published in the United States. We collected 1,023,084 PhD dissertation abstracts from the Proquest UMI database filed within the last thirty years, from 1980 through 2010. The dissertations are from 151 schools that have been classified as research-intensive by the Carnegie Foundation[3] in any of their three surveys of higher education conducted since 1994. These dissertations make up a reasonably large fraction of all PhD's granted by US academic institutions during that time span.

Each dissertation record in the Proquest UMI database contains a title, abstract, author, advisor, date, subject codes, and keywords. There are 263 subject codes in our dataset, which correspond to relatively high-level field designations (biochemistry, public administration, cultural anthropology, etc). These subject codes have been manually curated by taxonomists at Proquest, and have been updated over time to reflect changes in academia. On average, each dissertations contains 2.08 subject codes. Keywords are more tag-like, open-domain terms applied to by the dissertation author during filing. Abstracts contain an average of 179 non-stop words, corresponding to one to two paragraphs of text.

3 Methodology: modeling academic text

Our goal is to score and compare each academic institution by how oriented it is toward the future and toward the past. To do so, we first assume that each dissertation can be represented as a vector in some some vector space $V \in \mathbb{R}^n$ such that vectors satisfy two properties: composability and comparability. Vectors must be *composable* so that we can compute a signature vector of a set of documents from the vectors of the component documents, for instance by taking an average. Also, vectors must be *comparable* so that, for any two vectors $\vec{p}$ and $\vec{q}$, we can compute a robust similarity score $s(\vec{p}, \vec{q})$ of how alike are $\vec{p}$ and $\vec{q}$. Armed with these definitions, we can examine how well an individual school matches the future or past of academia as a whole.

Figure 1 (left) shows the history of academic dissertations as a year-to-year heat map computed using similarity vectors derived from a Labeled LDA [5] variant described below. We used the Stanford Topic Modeling Toolkit[4] to train the model. Each document's signature vector was taken as that document's distribution θ_d over latent topics. These signatures were averaged within each year, resulting in 30 signatures, one per year, each representing the expected distribution over topics for dissertations within that year. Signatures were compared using cosine similarity $\frac{\vec{p} \cdot \vec{q}}{|\vec{p}|_2 \cdot |q_2|}$ (which we found to be more stable and intuitive than other measures). Every signature was compared with every other signature, resulting in a dense matrix of year-to-year similarity within academia; each diagonal element is 1.0 (because every year compared with itself scores highly). Figure 1 demonstrates a property we hope and expect to find: that academia as a whole is moving and changing in a roughly consistent way over the course of our 30 year dataset.

We experimented with many ways of computing these signatures, including standard bag-of-word representations—l_2 normalized vectors of term counts (tf) or tf weighted by inverse document frequency (tf-idf)—as well as signatures based on Jaccard overlap in keywords or subject codes. We found these similarity scores were dominated by coding, transcription, and ontological migration issues inherent to the dataset. For example, the tf-idf year-to-year heatmap demonstrated a clear block structure showing that dissertations before 1987 all look alike; dissertations between 1987 to 1998 all look alike, and dissertations after 1999 all look alike. The reason for the observed block structure are largely historical: Proquest has been continually upgrading its data entry systems, internal representations, ontologies, and the like, over the course of time we see in our data. Therefore, systematic changes in the database align with systematic errors in the low-frequency terms, resulting the discovery of major "trends" that are, in reality, data artifacts.

[2] http://www.proquest.com/en-US/products/dissertations/

[3] http://classifications.carnegiefoundation.org/resources/

[4] http://nlp.stanford.edu/software/tmt/

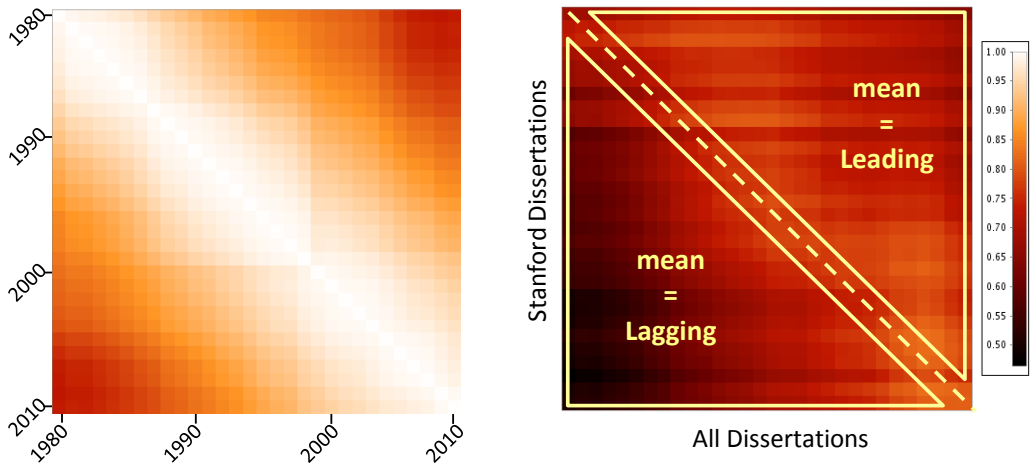

Figure 1: Left: the history of academia as seen through the lens of the topic model. Lighter is more similar. Right: Stanford's dissertations are (rows) are compared to dissertations from all schools (columns), so the scores above the diagonal represent Stanford's similarity to academia's future, or the extent to which it is leading. Similarly, the scores below the diagonal represent Stanford's similarity to academia's past.

As a result, the model we ultimately settled upon is based on a lower-dimensional representation inferred by a statistical topic models. Topic models have been used previously to study the academic history of computational linguistics [3] and to measure scholarly impact [2], among other applications in the social sciences. One drawback of latent topic models is that the topics themselves are often difficult to name or interpret. So in this work, we used a variant of Labeled LDA [5] that learned 8 latent topics associated with each of the 263 common subject codes, each describing some aspect of that subject code, and 8 common background latent topics, which capture non-field-specific academic language. Each document is then represented as a mixture of latent topics, each of which is itself associated with exactly one subject code. The final model contains 2,112 topics covering all broad areas of academia. Further quantitative study has shown that the relative rankings of schools are insensitive to the number of topics chosen per subject, and indeed, are comparable to the results inferred by vanilla LDA [1] models applied to the same documents, but the chosen model is more easily interpretable.

In addition to being able to compare each year to each other year, we'd like to be able to compare how much each individual institution looks like academia as a whole. To do so, we can compute a similar year-to-year heatmap, but this time with signatures from a particular school in a particular year anchoring the y-axis, and the signature for academia overall in a particular year. Figure 1 (right) illustrates one such school (Stanford). Note that in the upper-right triangle, we are comparing the signature of Stanford's dissertations to dissertations from all of academia that came in *later years*. Similarly, the lower-right triangle compares the signature of Stanford's dissertations to dissertations from all of academia that came in *earlier years.* We compute the *future similarity score* as the average similarity of school dissertations in a year to all dissertations in a later, which is equivalent to taking the mean of the values in the upper-right triangle. The *past similarity score* is computed analogously in the lower left.

4 Preliminary results

While this work is still in progress, our preliminary results show promise. Figure 2 presents the school-year to all-year similarity matrix for three more schools. On the left is another example of a school that tends to lead the future, Harvard. On the right is a mid-size university ranked near the bottom of the US News rankings in many fields, and we see that it tends to lag academia. In the center, we see a place where the model has failed: MIT has very few dissertations until 1990, at which point it began filing dissertations into the Proquest database at a rate comparable to its peer institutions. Unfortunately, this results in MIT's early years having very low similarity to everything else, thereby biasing our measure toward thinking that MIT is very past leaning (the dark stripe is on the top, and the rest is light). This is clearly an inappropriate conclusion - indeed, if we examine

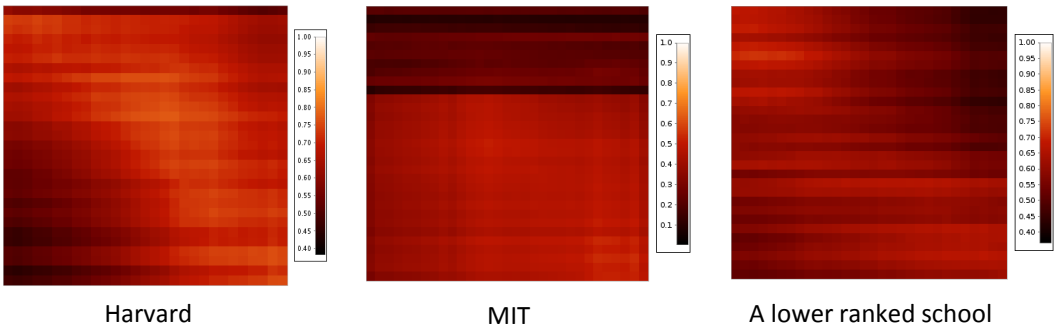

Figure 2: Examples of three univerisities' similarities to academia, by year. On the left, we see another university that has a higher leading score than lagging score. In the middle, we see a university whose results are skewed due to lack of data (MIT registered only a few dissertations per year with UMI until 1990). On the right, we see a university that tends to lag.

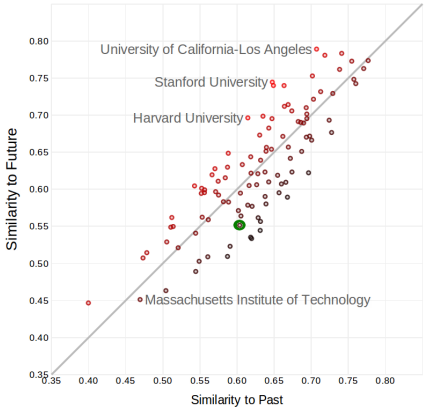

Figure 3: All universities plotted as dots, with similarity to the past on the x-axis and similarity to the future plotted on the y-axis. Dots over the diagonal line represent schools that are more future leaning, with dots under the diagonal line representing schools that are more past-leaning.

only the years after 1990, MIT is future-leaning - suggesting the need to automatically adapt our scoring system to only consider years that appear to contain enough valid data.

In order to explore our data, we built an interactive visualization on top of Protovis,[5] to examine all universities' similarity to the past and to the future on a single, interactive plot. A snapshot of that visualization is shown in Figure 3. The example schools shown so far are selected on the graph, with the poor performing school highlighted in green. Note that most schools on the top-left frontier are familiar, high-performing research universities, including many of the University of California schools, ivy league schools, and schools with good engineering departments. Most schools on the lower frontier, with the exception of schools such as MIT that have data sparsity issues, are less famous and lower ranked schools.

This extended abstract has only scratched the surface of the kinds of analysis we are performing on this unique dataset, and we have presented some of the challenges of drawing trustworthy, quantitative insight from the text dataset as it stands, with its messy history and complicated data irregularities. Nonetheless, we have found promising results via the use of topic models. Once these and remaining data artifact and trust issues have been worked through, we provide more specific analyses and ranking at the level of individual subject codes, and we plan to validate our findings beyond our own intuitions by way of comparing to existing rankings and expert intuitions.

[5] http://vis.stanford.edu/protovis/

Acknowledgments

This work was supported by the National Science Foundation under Grant No. 0835614. We thank ProQuest UMI for allowing access to the dissertation data and for providing insight into the dataset's history.

References

[1] D. M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *JMLR*, 2003.

[2] S. Gerrish and D. Blei. A language-based approach to measuring scholarly impact. In *ICML*, 2010.

[3] D.L.W. Hall, D. Jurafsky, and C.D. Manning. Studying the history of ideas using topic models. In *EMNLP*, 2008.

[4] J. P. Ostriker, C. V. Kuh, and J. A. Voytuk. A data-based assessment of research-doctorate programs in the united states. In *Committee to Assess Research-Doctorate Programs*. National Research Council, 2010.

[5] D. Ramage, D. Hall, R. Nallapati, , and C. D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-label corpora. In *EMNLP*, pages 248–256, 2009.