An analysis of data-oriented approaches to evaluating graduate schools in computer science

Benjamin Gafford

Abstract: There is significant debate over how to best rank graduate schools. Transparent, data-driven approaches have recently become more popular in the academic community as compared to the opaque, qualitative rankings provided by the *US News and World Report.* In this paper, we investigate how well these fledgling data-driven approaches match up with more qualitative approaches. We then leverage such data-driven approaches to forecast future trajectories of top graduate schools in computer science.

1. Introduction

Assessing and differentiating graduate schools is difficult. Many approaches have existed historically, starting with the more informal word-of-mouth reputation to the more formal qualitative approach taken by the *US News and World Report*, and more recently there has been a push to open metrics based analyses [**?**]. These open metric analyses leverage publicly available quantitative information such as citation counts, publications counts, and various publication impact metrics to provide greater clarity and reduce bias in ranking systems. Such approaches to rankings universities have risen in popularity in the field of computer science, and I believe more can be done with such data-driven approaches. With data-driven approaches, more interesting data analysis can be conducted, for example performing time-series analysis over longitudinal data to make forecasts about the future trajectory of a given institution. However, proponents of these approaches are quick to emphasize how these metrics-based approaches can fall short. Publications do not carry the same weight across sub-discipline and across venue, and even publications within a given venue can hold different intellectual value. Citation counts have similar problems. Additionally, when the metrics of evaluation are provided to the parties being evaluated, those parties will inevitably end up optimizing for such metrics through various means. Being liberal with citations and co-authorship among colleagues can be one easy way to artificially boost such metrics. Qualitative, report-based approaches are more robust to such tactics, however they can be opaque and biased towards existing high-prestige institutions. There is significant controversy around which ranking system to use. In this paper, I will resolve the following questions:

- **RQ1.** Do different metrics of evaluation capture different dimensions of success? Are some evaluation methods redundant to others? Can we identify interesting clusters when considering different metrics?
- **RQ2.** Which open source metrics most accurately fit *US News and World Report* rankings?
- **RQ3.** How has the productivity of top universities changed over time?

To investigate these research questions, I will combine datasets across different metrics and perform PCA analysis, clustering analysis, and create a model to predict institution rank. I will also perform a time-series analysis to make forecasts about the future productivity of top universities using available data-driven metrics.

2. Methods

The data used in this paper comes from a number of sources. The raw publication data from 1970-2019 comes from DBLP, which has been used in conjunction with CSRankings [2] data. This publication data covers 435 institutions worldwide. CSRankings data also includes crowd-sourced information regarding the current affiliation of professors, and semantic information about conferences. This semantic information includes the related sub-discipline and sub-sub-discipline for each conference, and notes the "top" conferences in each area. From the National Science Foundation, we used a dataset that contained the NSF Graduate Research Fellowship (NSFGRFP) award information from 2000-2012 [3]. We used citation data and *US News and World Report* data from [4], and they scraped citation data and from Google Scholar.

To merge these datasets together, university name was used. One difficulty to doing this is the inconsistency of university name across datasets. For example, “University of California – Berkeley” could be referred to as “UC Berkeley”, “University of California - Berkeley”, or “University of California Berkeley”, and many more variations therein. We employed fuzzy string matching and several text mutations to gain consistency across datasets. Since we were considering institutions globally, we needed to fix missing data points for comparable foreign institutions. Relevant missing information includes the *US News and World Report* ranking and the number of NSFGRFP recipients. To do so, we used K nearest neighbors imputation with a weighted average to compute these missing values. Because there is substantial publication data for all of these institutions, and only several global universities in the examined set, I believe this is the best option available and does not compromise the data significantly.

I also included number of authors for each institution, as measured by the number of authors with at least one publication, and the mean number of authors per paper in each venue to hopefully shed some more light on relationships between various institutions.

The data was further consolidated to only include institutions with greater than 273 total publications (leaving only 100 institutions remaining). Institutions with fewer publications were generally less interesting, more variable, and occluded the relationships between the higher-fidelity data. Data-driven approaches are only viable when there is sufficient data available.

Specific data-driven metrics that will be included in my analyses include the following:

- NSF GRFP recipients planning to go to X school
- Number of publications in top tier venues
- Number of citations

2.1. Principal component analysis (PCA) and clustering

Since this dataset contained many variables, PCA was used to easily drop dimensions and find clusters in data. Clustering was done using partitioning around medoids (PAM) clustering.

2.2. Modeling

We used LASSO regression to create a model to accurately predict the *US News and World Report* ranking given the variables of interest in the model. LASSO handles for multicollinearity, and can identify the variables that matter. In this case, we are more interested in identifying which variables provide new information and which variables are largely redundant to one another.

2.3. Forecasting time series groups

We leveraged longitudinal publication data to forecast future publication data. As a reminder, publication data is one of the popular data-driven metrics. To forecast, we used an Exponential Smoothing model because there was no clear pattern in the data, and we wanted to avoid spurious bumps and lumps in our predictions.

3. Results

RQ1. Do different metrics of evaluation capture different dimensions of success? Are some evaluation methods redundant to others? Can we identify interesting clusters when considering different metrics?

When applying PCA as a means of dimension reduction to to the top 100 schools, we can see that these data-driven metrics appear to be highly collinear with one another, and with the qualitative *US News and World Report* rankings. Looking at the PCA loadings over the top 100 schools in A.1, A.2, and A.3, we can see that different metrics all result in roughly the same effect in the top 3 dimensions. From examining the loadings further, we can see that dimension one can be interpreted as the overall excellence of the school, and dimension two can be interpreted as the degree of collaboration in papers, where lower values correspond to higher average numbers of co-authors on papers. It is important to note here that higher amounts of co-authors could be a result of high collaboration, of more undergraduate projects (undergraduate projects generally involve higher number of co-authors due to , or of groups of individuals seeking to artificially boost their rankings through liberal co-authorship. Clustering analysis primarily groups schools according to their position along the “excellence” dimension, as shown in A.4.

To investigate further these various metrics in PCA and clustering, we focused on the top 10 schools according to publication count. From this clustering, we found 3 interesting clusters among institutions, as shown in A.7.

We found that dimension 1 remained the same as a proxy for excellence, however dimension 2 now reflected the relative excellence of the university outside of computer science according to the *US News and World Report* ranking, and the relative size of the school. A worse overall rated school and a higher author count will correspond to a lower value on the 2nd dimension. The loadings for these dimensions can be seen clearly in A.8.

RQ2. Which open source metrics most accurately fit *US News and World Report* rankings?

To determine which metrics most accurately predict the *US News and World Report* ranking, we chose to use LASSO modeling as described in section 2.2. We found that the most relevant variables were publications to the Theory CS sub-discipline and to total citations, as seen in A.5. With just these two values, our model had an RMSE of 0.697, which indicates that prediction errors were relatively small on average, as each unit increase corresponds to a ranking that is reduced by one (out of 100). This can be seen in A.6.

RQ3. How has the productivity of top universities changed over time? The forecasting in A.9 shows several things. For one, it shows that in recent years the growth in publications from year to year has diminished, and most graduate schools have stagnated with respect to their number of publications, even seeing a dip around 2017. Since only top venues “count” in this equation, and the top venues have remained largely the same over the past couple of years in terms of number of papers they accept, it would make sense that these top schools would struggle to increase output to these conferences. For one school to increase output, others will necessarily need to decrease output. The only school with a notable upward trend in publication count is UC Berkeley, so it could potentially rise in the ranks in the coming years. For the residuals for this time-series analysis, refer to A.10 in the appendix.

4. Discussion

These results suggest the qualitative and quantitative evaluation techniques addressed in this paper are functionally very similar. This robustness of ranking despite different metrics suggests that the rankings are likely accurate.

There are several limitations to this study. For one, much of the publication data is crowd-sourced, and there is missing information regarding authorship and affiliation, which leads to a lot of missing and therefore discounted data points. Since this crowd-sourcing campaign originated among professors at top CS graduate schools in the United States, it is likely that individuals outside of this area of influence are the ones to have missing data. This will disproportionately harm foreign institutions, and smaller institutions in the United States.

Additionally, different datasets were looking at different slices of time with respect to the schools. The citation data only included total citation count, and not by year, so publication data therefore was used the same way in modeling and clustering analysis. *US News and World Report* data, however, was from a single point in time rather than the summation of all historical contributions. Among high-ranked institutions, at least, this will likely have little effect, as we can see the relative rankings of these institutions has remained rather stagnant over the past 40 years. For lower ranked institutions, this will likely have a larger effect as there will be greater variability in performance from year-to-year due to smaller faculty sizes. Theoretically, longitudinal data is available from all of these different data sources. One potential avenue for future work would be to scrape such data from the relevant sources, clean it, and make it available for public use. Then, researchers could do time series analysis across more variables, and could also include more recent information in the open metric based data.

Related work includes that done by the CSMetrics team [1]. They also used Microsoft Academic Search to gain information regarding citation networks, citation counts, and active affiliations. We considered using these sources, however they entailed a great deal of scraping beyond what should be done in R, and also external barriers including API keys and dealt with data too big to be stored and processed on an old laptop.

References

[1] CSMetrics. Csmetrics data, 2019. csmetrics.org.

[2] CSRankings. Csrankings data, 2019. csrankings.org.

[3] data.gov. Nsf grfp recipients 2000-2012, 2019. data retrieved from data.gov, https://inventory.data.gov/dataset/43c096b5-1508-4c34-84de-f9c3dd4f8ba5/resource/1d9c84a4-17ac-4f51-b488-d661a6c0b408/download/userssharedsdfnsfgrfpawardeesandhonorablementions20002012.csv.

[4] Slobodan Vucetic, Ashis Kumar Chanda, Shanshan Zhang, Tian Bai, and Aniruddha Maiti. Faculty citation measures are highly correlated with peer assessment of computer science doctoral programs, 2017. data retrieved from, http://www.dabi.temple.edu/ vucetic/CSranking/details/.

Appendix A. Figures

Figure A.1: PCA dimension 1 contributions for top 100 schools.

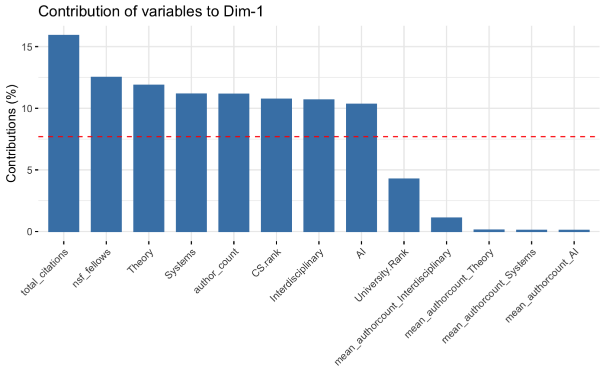


Figure A.2: PCA dimension 2 contributions for top 100 schools.

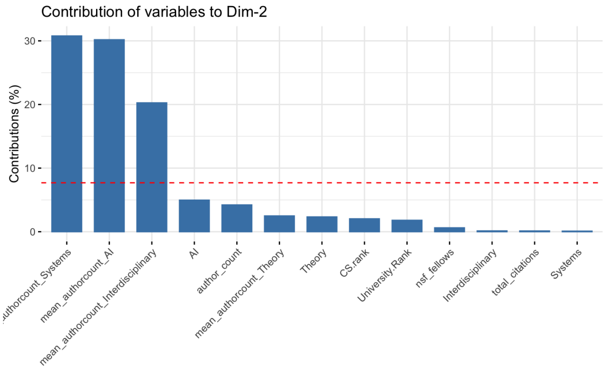


Figure A.3: PCA dimension 3 contributions for top 100 schools.

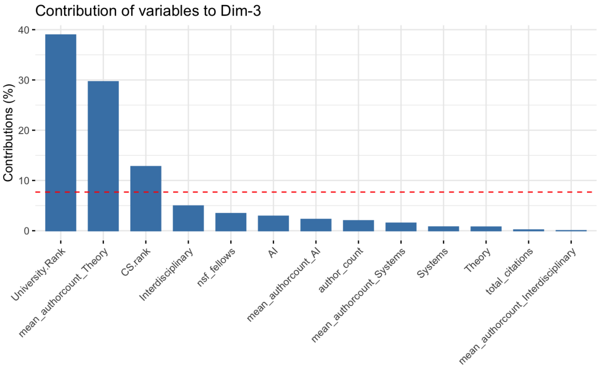


Figure A.4: PAM clustering over top 100 schools.

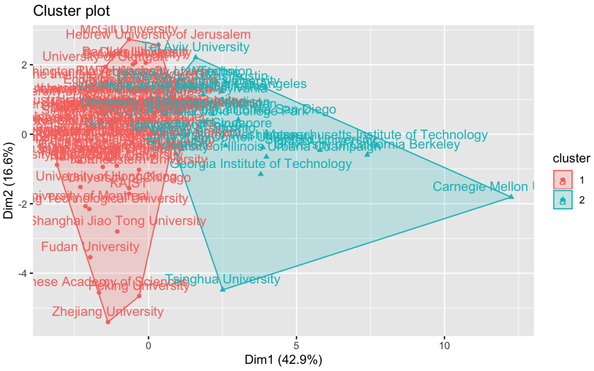
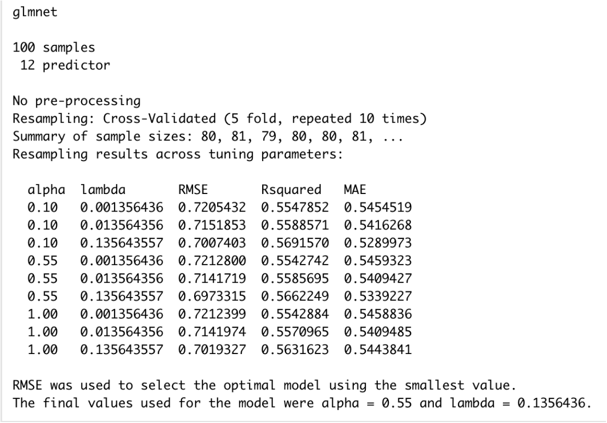


Figure A.5: Lasso model variables to predict US News ranking.

```
12 x 1 sparse Matrix of class "dgCMatrix"
                                               1
(Intercept)                         7.093211e-17
AI                                  .
Interdisciplinary                   .
Systems                             .
Theory                             -2.730116e-04
mean_authorcount_AI                 .
mean_authorcount_Interdisciplinary  .
mean_authorcount_Systems            .
mean_authorcount_Theory             .
author_count                        .
nsf_fellows                         .
total_citations                    -3.205766e-02
```

Figure A.6: Lasso model evaluation to predict US News ranking.



```
glmnet

100 samples
 12 predictor

No pre-processing
Resampling: Cross-Validated (5 fold, repeated 10 times)
Summary of sample sizes: 80, 81, 79, 80, 80, 81, ...
Resampling results across tuning parameters:

  alpha  lambda       RMSE       Rsquared   MAE
  0.10   0.001356436  0.7205432  0.5547852  0.5454519
  0.10   0.013564356  0.7151853  0.5588571  0.5416268
  0.10   0.135643557  0.7007403  0.5691570  0.5289973
  0.55   0.001356436  0.7212800  0.5542742  0.5459323
  0.55   0.013564356  0.7141719  0.5585695  0.5409427
  0.55   0.135643557  0.6973315  0.5662249  0.5339227
  1.00   0.001356436  0.7212399  0.5542884  0.5458836
  1.00   0.013564356  0.7141974  0.5570965  0.5409485
  1.00   0.135643557  0.7019327  0.5631623  0.5443841

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were alpha = 0.55 and lambda = 0.1356436.
```

Figure A.7: PAM clustering over top 10 schools.

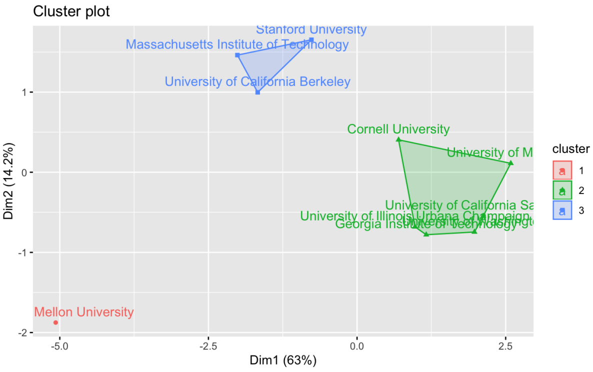


Figure A.8: PCA analysis over top 10 schools.

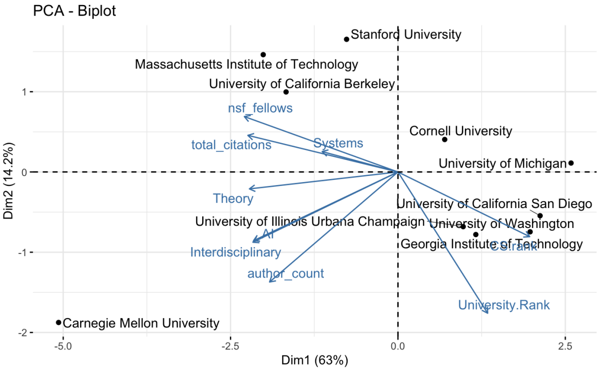


Figure A.9: ETS model forecasts for the top 10 CS schools.

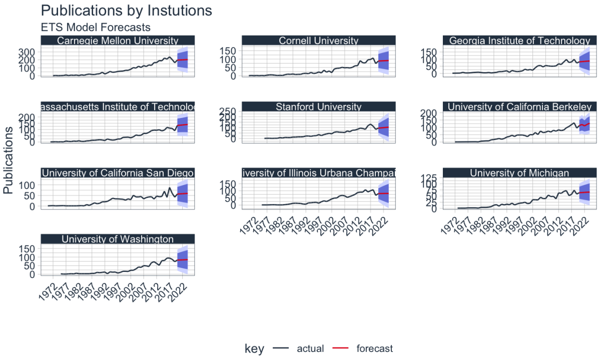
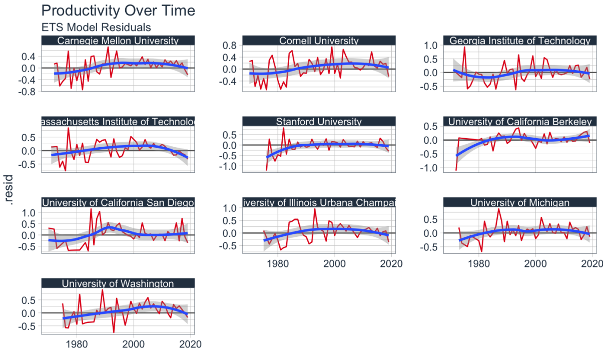


Figure A.10: ETS model residuals for the top 10 CS schools.


Productivity Over Time
ETS Model Residuals
Carnegie Mellon University
Cornell University
Georgia Institute of Technology
assachusetts Institute of Technolo
Stanford University
University of California Berkeley
University of California San Diego
iversity of Illinois Urbana Champa
University of Michigan
University of Washington
.resid