A quantitative criticism of the 2010 NRC assessment of graduate programs in Mathematics

Peter J. Mucha
Department of Mathematics, University of North Carolina at Chapel Hill

This document presents a modest visual exploration of the data from the 2010 National Research Council (NRC) assessment, focusing here on the field of Mathematics. By so doing, we aim to demonstrate the limited nature of the resulting rankings. By further documenting errors in those variables that influence the rankings, we identify severe problems with the results. Combining this investigation with other critiques of the NRC assessment, we argue that it would be a serious error for the American Mathematical Society to implicitly endorse the 2010 NRC rankings by using them to re-group the Departments of Mathematics.

When the 2010 NRC rankings were released, the departments and the universities who did well quite understandably and naturally trumpeted their good standing. At the same time, however, a significant chorus quickly voiced serious concerns with multiple fundamental aspects of the new rankings. Of course, every ranking system has its flaws; but some of the criticisms of the 2010 NRC rankings go to the very foundation of the data gathering and statistical methodology employed, explicitly and directly undermining the central "data-based" claims that the rankings were purportedly created to address. Even the NRC itself fails to recommend these rankings as anything more than "illustrative" [1]:

> *The committee had originally planned to combine the R- and the S- rankings into a single range of rankings, which is the approach outlined in the 2009 Methodology Guide. The production of rankings from measures of quantitative data turned out to be more complicated and to have greater uncertainty than originally thought. As a consequence, the committee did not combine the two measures, and instead has presented them as two illustrative rankings. Neither one is endorsed or recommended by the National Research Council as an authoritative conclusion about the relative quality of doctoral programs.*

Similarly, the COO of the National Research Council and the chair of the Committee on an Assessment of Research-Doctorate Programs together wrote [6]

> *We want to stress that the rankings provided in the NRC report are really intended to be "illustrative." They are not intended to be definitive; they are not endorsed by the NRC above other alternatives that might be constructed. Instead, they are examples of two ways of deriving faculty values for determining weights for making rankings, which illustrate how stakeholders can apply their own weights.*

In Mathematics, the current "Groups" structure of mathematics departments relies on the 1995 NRC rankings. With the current Groups now based on ever-aging data, one might reasonably propose by default that new groups of Mathematics departments be assigned based on these 2010 NRC rankings. This document aims to present a case that it would be a mistake to compound

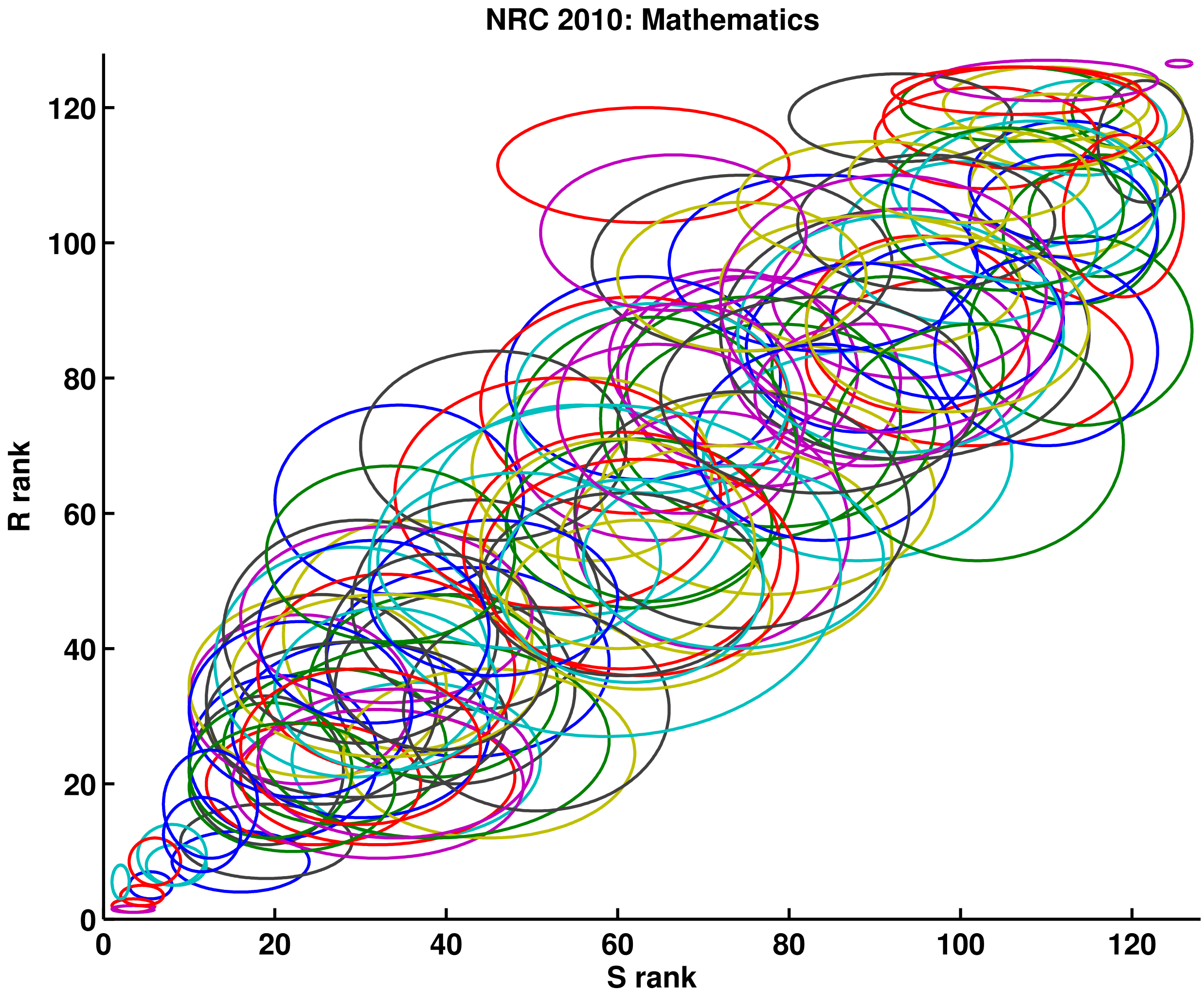


Figure 1: Comparison of R and S ranks in the 2010 NRC Mathematics ranking. Each ellipse represents the ranks for a single school, spanning the published 5th and 95th percentiles in both R and S. This figure is inspired by similar images created by Bloomfield & Stefanski [2], but here uses the revised data released in April 2011. The large sizes of the ellipses emphasize the wide uncertainty present in the rankings, except for a small number of departments appearing in the lower left corner of the figure.

the flaws in the 2010 NRC rankings by using them for anything so official and of such lasting importance as a regrouping of the Mathematics departments.

We presuppose here that the reader has already been exposed to and has some understanding both of the intent of the 2010 NRC rankings and of many of the criticisms written by others (including, e.g., [4, 5, 7, 9, 10, 11, 12, 14]). We do not recapitulate here the suite of excellent arguments already made about the general flaws in the 2010 NRC methodology and results that have been identified broadly across disciplines. We instead focus our critique here on the 2010 NRC data specific to Mathematics, including visual display of that data and significant questions about the validity of those variables that strongly influence the rankings in Mathematics.

We start in Figures 1 and 2 with simple visual depictions of the two “illustrative” rankings in Mathematics in the 2010 NRC assessment, using the “R” (Regression) and “S” (Survey-based) coefficients on the 20 variables used. As is evident in the figures, the reported 5th-to-95th percentile

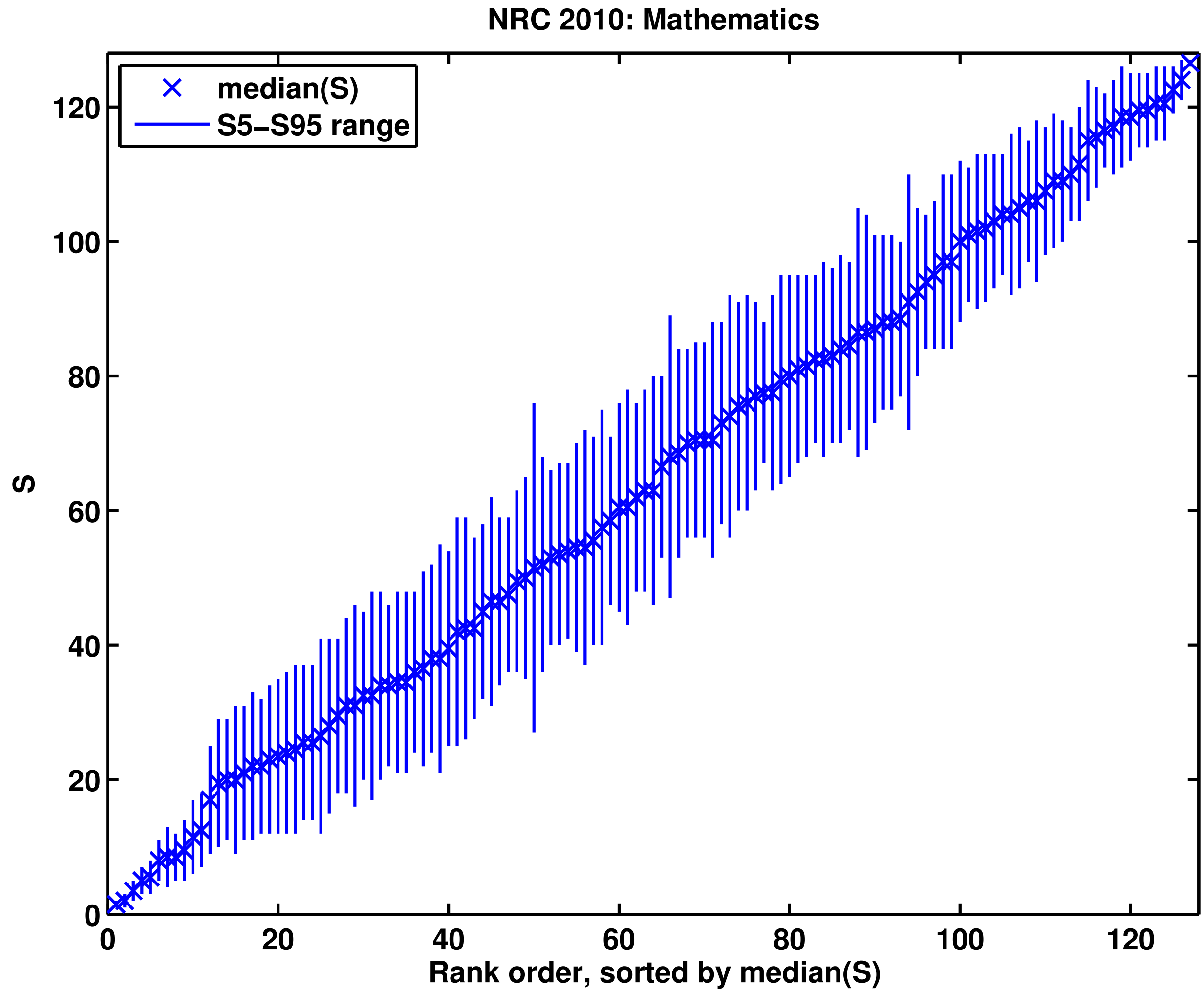


Figure 2: Survey-based "S" ranks v. rank order obtained by sorting the medians of the reported 5th and 95th percentiles in Mathematics for each university. Apart from a small number of departments represented in the lower left corner of the figure, there is large uncertainty represented in the published confidence interval range (represented here by vertical bars).

confidence intervals are very wide. We additionally remark that there are no clearly identifiable "tiers" of departments given the large amount of uncertainty indicated by these wide ranges, except perhaps the small jump visually identifiable in Figure 2 after the "top 11" departments.

As we continue to visualize the 2010 NRC rankings in Mathematics, it is important to understand that these rankings are based on models that are linear in the standardized program values. That is, the 2010 NRC methodology converts every program value into a z-score, quantifying the number of standard deviations away from the mean for that program value relative to those reported by other programs in the same field. These z-scores are then multiplied by coefficients (drawn from regressions for the R rankings and from surveys for the S rankings) to quantify the quality of each program. We proceed here in our focus on Mathematics by direct examination of the z-scores and the coefficients that multiply those z-scores in the rankings.

Starting with the R rankings, one might ask whether the "R" stands for "Reprehensible" instead of "Regression." The regression coefficients obtained in Mathematics are, in a word, offensive. If anyone wants to use this "illustrative" data as a roadmap to improve the R ranking of a Mathematics

To-Do List to Improve* Mathematics Departments *according to 2010 NRC R rankings				
		Regression Coeffs.		
"To-Do Item"	Variable	−1 SD		+1 SD
Reduce % Non-Asian Minority Faculty	V5	-0.024	to	0.002
Reduce % Female Faculty	V6	-0.008	to	0.015†
Reduce % fully supported 1st yr. students	V9	-0.031	to	-0.013
Discourage external funding of 1st yr. students	V10	-0.019	to	0.004
Reduce % Non-Asian Minority Students	V11	-0.024	to	-0.002
Reduce % Female Students	V12	-0.029	to	-0.008
Extend time to degree	V16	0.039	to	0.050

Table 1: One might propose using the variables and corresponding regression coefficients from the 2010 NRC assessment of Mathematics programs to generate this to-do list of ways to "improve" a department. *To be clear, this proposed use is definitively absurd; but anyone effectively endorsing the 2010 NRC rankings by its official use should ask whether they are willing to green-light this kind of use too.* †: In the original September 2010 release, the regression coefficients for V6 were published as −0.045 to 0.008, so this item remains here for anyone who wants to take a "better safe than sorry" approach when "improving" their Mathematics program.

department, the to-do list is presented in Table 1 (complete with the regression coefficients that multiply the standardized program values). Speaking personally for a moment, it's shocking that anyone looked at these regression coefficients and thought this was a reasonable table to publish; these numbers by themselves should be sufficient grounds to throw out any official use of the 2010 NRC rankings altogether. Others might try to counter with "We'll only use the S rankings," but such arguments completely ignore the larger lessons offered by these "Reprehensible" R rankings. The variables measured to purportedly quantify the perceived qualities of mathematics graduate programs are clearly insufficient to the task if they instead identified the sad realities of the low numbers of females and minorities in some of the top departments. Conveniently turning a blind eye to these offensive R results ignores the fundamental problems of how little data appears in the assessment at all and the even smaller amount of that data that might in any way describe the quality of an academic program. *Indeed, the offensiveness of the R rankings in Mathematics is "illustrative" of the fundamental foundational flaws underlying the 2010 NRC assessment.*

While one might reasonably find the above information to be more than sufficient reason to steer clear of the 2010 NRC rankings altogether, for completeness we continue here with some visualizations of the S rankings. Figure 3 presents the standardized program values in Mathematics plotted versus the median of the 5th and 95th percentile S rankings for each university. It is visually apparent in the figure that most of the variables have little or no correlation with the S rankings. That is, while it sounds impressive to say that the spreadsheet containing the complete 2010 NRC data is 35MB, it would seem that only 6 variables impact each institution's ranking in Mathematics. For those keeping score, storing both the program values and z-scores as 16-bit Half variables[1] across 129 sets[2] of all 20 variables requires just over 10kB. And as noted above, probably only 30% of these 20 variables substantively matter in the S rankings in Mathematics.

Looking more closely at the direct contributions of the variables to the S rankings, we plot in Figure 4 the products of the standardized values and survey-based coefficients (taking the median

[1] Given the discussion about V1 below, even this much precision is questioned.

[2] There are 127 programs listed in Mathematics, plus the R and S ranking coefficients (which only multiply the z-scores but appear as a range of ± 1 SD coefficients), equivalent to 129 sets of values, each requiring 80 bytes, ignoring the small multiple imposed by the different corrections for variation over time included in the spreadsheet.

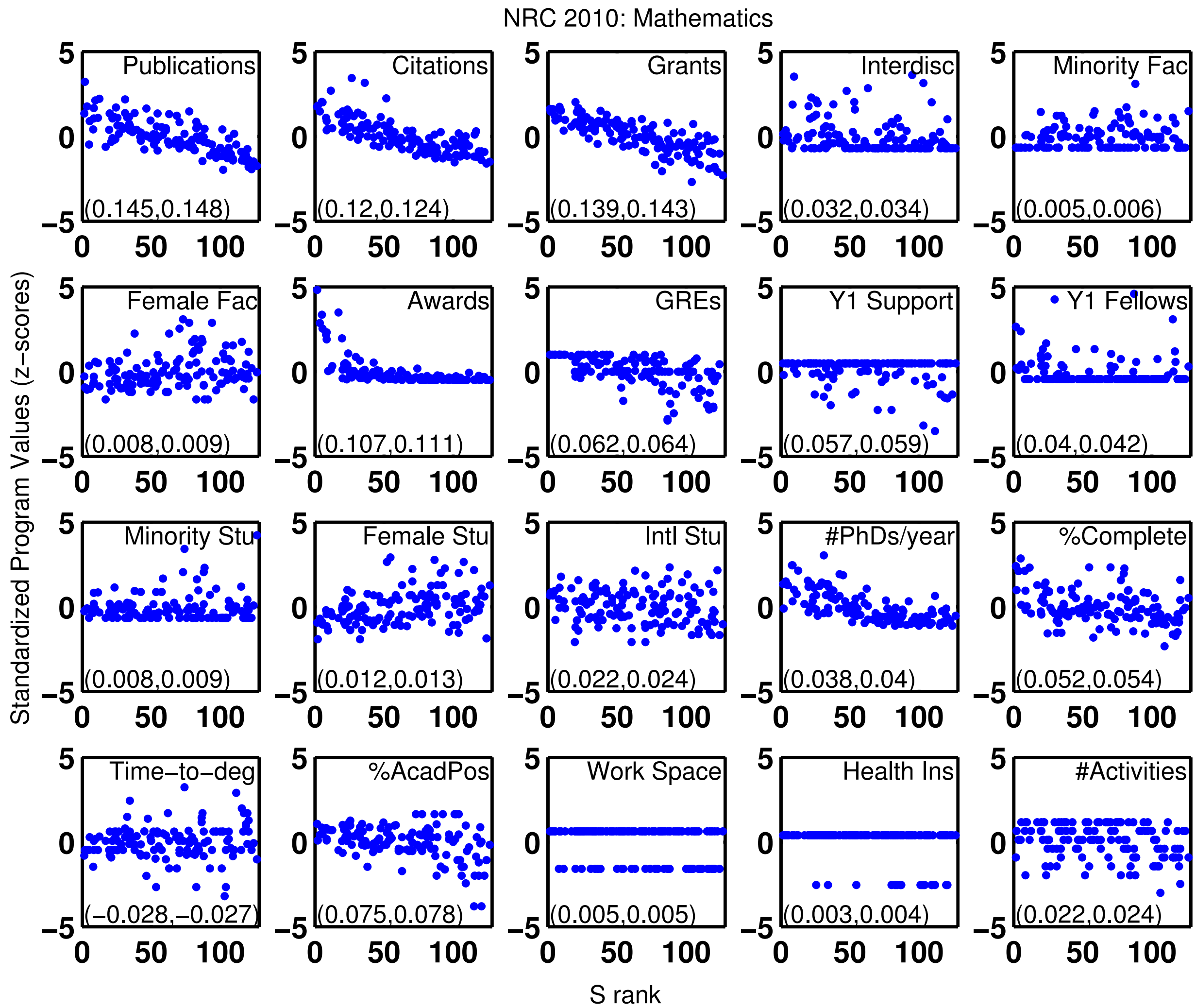


Figure 3: Standardized program values for each of the 20 variables used in the 2010 NRC rankings, plotted versus the median of the 5th and 95th percentiles for the Survey-based "S" ranking in Mathematics of each university. This visualization suggests that the variables most strongly correlated with the S ranking in Mathematics are V1: Publications per Allocated Faculty, V2: Cites per Publication, V3: Percent Faculty with Grants, V7: Awards per Allocated Faculty, V8: Average GRE, and V14: Average PhDs 2002 to 2006.

of each range) for the programs whose median S value places them in the #25–75 range (using the median S values to sort results). Again, it would seem that the contributions of only a small fraction of these variables correlates with the S ranking of the departments. Moreover, looking at the spread of these variables begs questions about the level of the noise in the results, which is of course borne out in the uncertainties communicated by the large ranges between the 5th and 95th percentiles in the published rankings (as visualized in Figure 2). That is, the wide spreads between these percentiles are intimately related to the spread of the underlying data points plotted in Figures 3 and 4. Perhaps more troubling to any individual institution appearing as an outlier in any of these plots is the question of the impact of that single variable on their overall ranking, particularly where that variable does not appear to be well correlated with the rankings. As just one for instance, given the absence of correlation with V17: Percent Students in Academic Positions (and given the likely noise in the measurement of that variable), the rewarding and punishing of

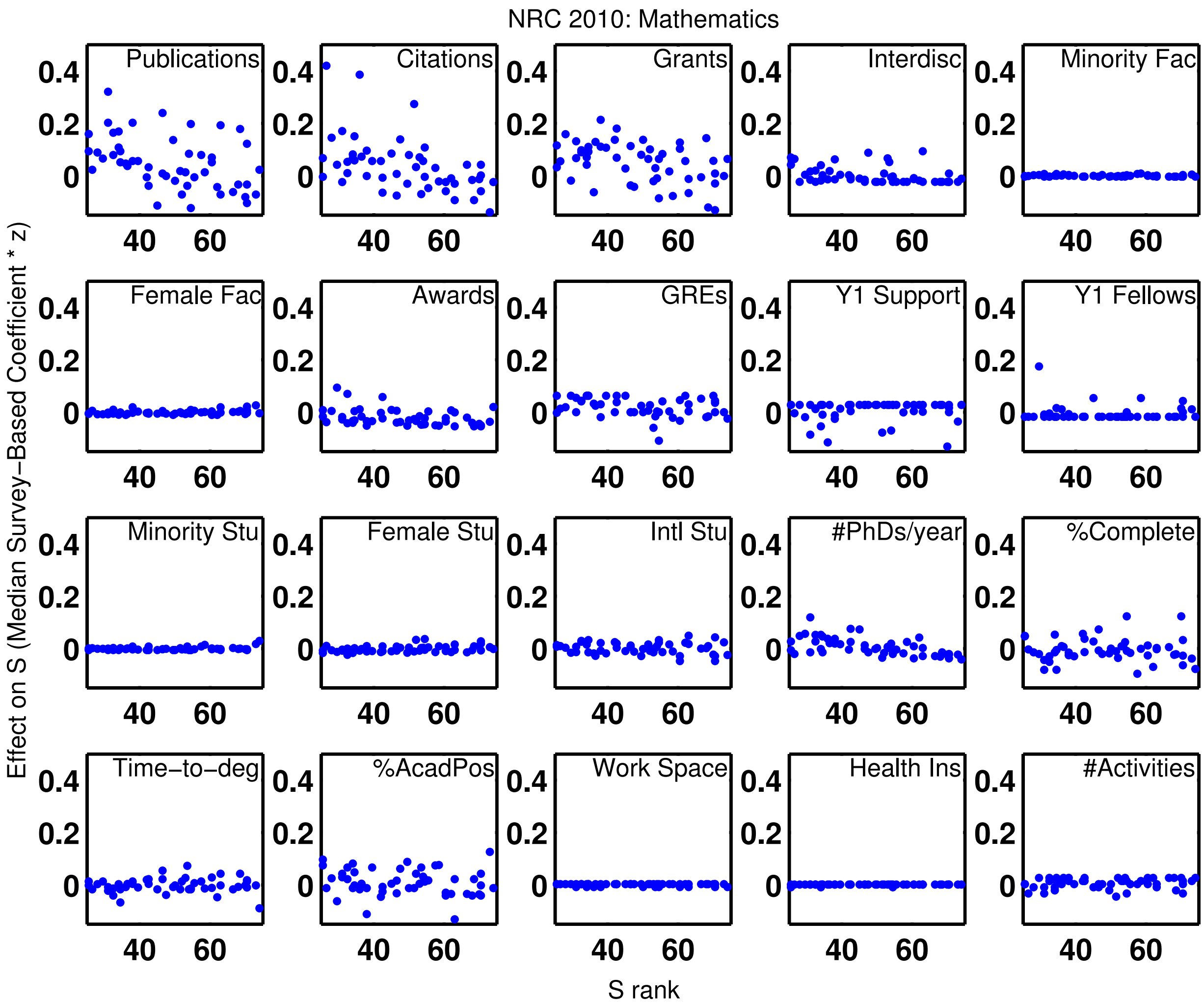


Figure 4: Contribution to S ranking score (standardized program values multiplied by the median S coefficients) for each of the 20 variables used in the 2010 NRC rankings, plotted versus the median of the 5th and 95th percentiles for the Survey-based "S" ranking in Mathematics of each university. This visualization suggests that the only variables substantively impacting the trends in the S ranking in Mathematics are V1: Publications per Allocated Faculty, V2: Cites per Publication, V3: Percent Faculty with Grants. There appears to be a much smaller correlation with V7: Awards per Allocated Faculty and V14: Average PhDs 2002 to 2006.

departments based on this variable would seem highly questionable.

In the spirit of completeness (or complete dismantling) that permeates the present essay, let's proceed by asking a few more questions about the variables that appear to matter in the 2010 NRC S ranking in Mathematics.

- V1: Publications per Allocated Faculty would seem to be non-controversial enough as a variable by which to measure a program, if they were actually being counted correctly. However, as noted in [4], there are a number of general problems with both the selection and the measurement of this and other variables. Most notable for Mathematics is that the publication variable can have significant miscounting due to incorrect matching of the faculty names on the rosters of each program. The author being of course most familiar with his own program, we can quickly and easily identify nearly *twice* as many publications as that identified by the

NRC. Using precisely the same source (ISI, 2000-2006), and using the same faculty roster, the sum of faculty publications totaled 340.5 papers (after rescaling with the same percentage allocation as on the faculty roster), corresponding to 1.424 publications per allocated faculty per year, while the NRC spreadsheet lists 0.725. For comparison, the standard deviation of V1 in Mathematics is only 0.310. That is, the measurement error on this single data point appears to be 2.25 times larger than the standard deviation in the observed program values! Importantly, this variable makes some of the largest contributions in the S rankings and it was one of the variables determined by the NRC, not provided by the departments. After identifying this error in the measurement of Mathematics and similar errors in other fields at UNC-CH, the Deans of the Graduate School at UNC-CH sent a formal error response memo to the NRC requesting clarification; but they never received a response to these queries.

- V2: Cites per Publication would similarly seem to be non-controversial; but if the NRC cannot correctly identify half of a department's publications, are we supposed to believe that this citation count is correct? Like V1, this data point was obtained by the NRC itself, not from the departments. As yet another example of flaws in the data, the original September 2010 data release completely omitted publications from 2002 [4].

- V3: Percent Faculty with Grants is a reasonable enough variable, given that faculty having grants is generally recognized as a good thing. Do you know what might be more interesting to prospective graduate students trying to figure out how much they'll have to teach and how much they might be supported off research awards? The total numbers of grants per faculty. Or the total dollar figure of those grants. But neither of those values appear in the 2010 NRC assessment.

- V7: Awards per Allocated Faculty is also seemingly non-controversial on its face. But the definition of what constitutes an award was apparently so uncertain that correcting the list of awards was one of the reasons for the April 2011 release of corrected data [9]. It would seem that despite years of careful consideration leading up to the September 2010 data release, the list was then changed in less than 7 months for re-release. Again taking a special interest in Mathematics at UNC-CH, the number of awards increased more than fourfold between these two releases. This variable is of course also suspect to name matching problems [4].

- V8: Average GRE scores of enrolled students appear to correlate with the rankings in Figure 3, but appear to be making very little direct contribution in Figure 4. Nevertheless, it is important to stress that one of the primary stated goals of the 2010 NRC assessment was to get away from measures based on reputation. So instead of querying the faculty expertise about the quality of programs in their respective fields, these rankings indirectly query the prospective graduate students about their perceived notions of the reputations of programs, while simultaneously ranking those graduate students based only on the GRE scores they obtained prior to entering the program that is being measured. Except for as an indirect and flawed measure of the quality of the student body, the quality of the graduate education offered by a program is not measured here. Instead, this variable primarily but indirectly measures reputation—a more direct measurement of reputation would seem to be preferable.

- V14: Average PhDs 2002 to 2006 directly rewards larger programs over smaller programs. One can certainly make some valid arguments about the potential importance of programmatic breadth. But beyond some size, does it really matter how many PhD students a program graduates per year? The meaning of this variable is also skewed by averaging only non-zero years [4].

Conclusions

I admit clearly and unabashedly that I have a natural motivation to passionately press the argument against any official use of the 2010 NRC assessment, given how poorly my department fared in the rankings. But I am equally motivated here on grounds of intellectual honesty. Indeed, I believe I have served as a strong voice on my campus drawing attention to the ways in which these rankings are especially flawed, and I honestly believe that my credentials as a mathematical scientist and my approach to this issue have helped some people on my campus understand that they should take these rankings with a salt-lick-sized grain of salt.

At least two other academic societies have already spoken out against the 2010 NRC assessment (the American Sociological Association [7] and the Computing Research Association [12]). Given the numerous problems identified above and elsewhere, it would be disgraceful for the AMS, ASA, MAA and SIAM to implicitly endorse these rankings by using them to regroup the departments. Such endorsement would send completely the wrong message to other disciplines about the quality of the methodology and the accuracy of the data used therein. As mathematical scientists, we should not implicitly endorse these rankings in any way; rather, we should have been the ones out front calling them inappropriate.

References

[1] *A data-based assessment of research-doctorate programs in the United States, Frequently Asked Questions*, `http://sites.nationalacademies.org/pga/resdoc/pga_051962`.

[2] P. Bloomfield and L. Stefanski, *R programs for visualizing the NRC rankings data*, `http://www4.stat.ncsu.edu/~stefansk/NRC_R_Programs/`, (2010).

[3] V. Burris, *The academic caste system: Prestige hierarchies in PhD exchange networks*, American Sociological Review, 69 (2004), pp. 239–264.

[4] J. Carpenter-Hubin and L. McClelland, *Data comparability, utility, and collection methods of the 2010 nrc assessment of doctorate programs: Limitations and suggestions for improvement*, Presented at "Assessing Research-Doctorate Programs: A Convocation on Analytic Uses and Future Directions", (2011).

[5] J. R. Cole, *Too big to fail*, The Chronicle of Higher Education, (April 24, 2011).

[6] E. W. Colglazier and J. P. Ostriker, *Counterpoint: Doctoral-program rankings—the NRC responds*, The Chronicle of Higher Education, (October 17, 2010).

[7] J. H. Evans, S. T. Hillsman, A. Perrin, M. L. Small, and S. S. Smith, *Report to the American Sociological Association Council regarding the 2010 National Research Council assessment of doctorate programs*, `http://www.asanet.org/images/press/docs/pdf/2010_NRC_Rankings_of_Doctorate_Programs.pdf`.

[8] J. H. Fowler, B. Grofman, and N. Masuoka, *Social networks in political science: Hiring and placement of Ph.Ds, 19602002*, PS: Political Science and Politics, 40 (2007), p. 729739.

[9] D. Glenn, *NRC plans to release revised doctoral-program rankings soon*, The Chronicle of Higher Education, (March 6, 2011).

[10] ——, *Computer scientists cry foul over data problems in NRC rankings*, The Chronicle of Higher Education, (October 6, 2010).

[11] ——, *A critic sees deep problems in the doctoral rankings*, The Chronicle of Higher Education, (September 30, 2010).

[12] E. Grimson, *Ratings redux, on behalf of the CRA board*, `http://www.cra.org/resources/crn-archive-view-detail/ratings_redux/`.

[13] S.-K. Han, *Tribal regimes in academia: a comparative analysis of market structure across disciplines*, Social Networks, 25 (2003), pp. 251–280.

[14] S. Stigler, *The NRC rankings: Further thoughts from Stephen Stigler*, http://chronicle.com/blogs/measuring/the-nrc-rankings-further-thoughts-from-stephen-stigler/27511, (2010).

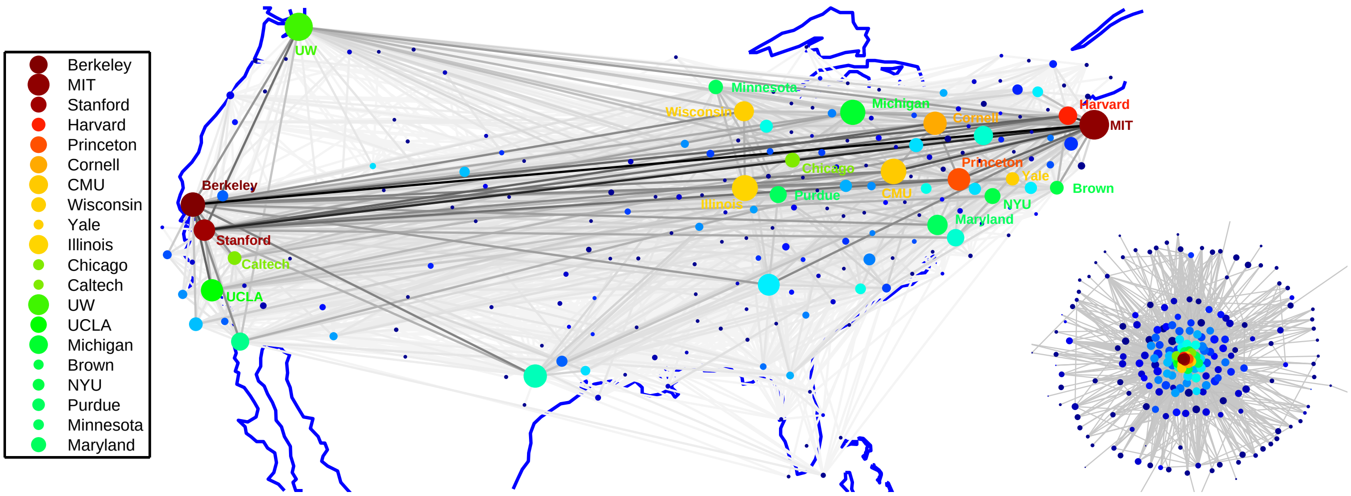
Appendix: Mathematical Genealogy & Department Prestige

As an appendix to the data-based discussion in the main part of this document, we include here for completeness an alternative data-based ranking in terms of centrality in the Ph.D. exchange network linking doctoral training in Mathematics across universities in the United States.

The attached 1-page preprint has been submitted to *Chaos* to appear as one of the winners of the 2011 Gallery of Nonlinear Images (http://www.aps.org/units/gsnp/gallery/index.cfm). Inspired by previous use of network theory to investigate prestige hierarchies in different academic fields (see, e.g., [3, 8, 13]), we investigated the directed network of universities in the United States counting the numbers of doctoral students from each university who later advise students at another university, as specified in the data from the Mathematics Genealogy Project (http://www.genealogy.ams.org/) for the period 1973–2010. The attached manuscript admittedly only represents a preliminary analysis of this data, without any dynamic variability. Moreover, the use of MGP data for this purpose in isolation raises a variety of questions. One might ask about the effect of the fact that a variety of students from other departments closely aligned with the mathematical sciences are listed in the MGP. There may also be effects induced by the overall sizes of each program in this network representation.

Despite the necessary and reasonable questions about this approach, the resulting figure strikingly indicates that the authority scores calculated on this network correlate with the rankings from U.S. News & World Report and the 1995 NRC rankings (higher network authority corresponds well with lower rank numbers), particularly near the upper end of the prestige hierarchy. These results are qualitatively consistent with previous work: in [3], 82–84% of the variance in the 1995 NRC rankings is captured by a model in terms of social capital (another measure of centrality) alone in the Ph.D. exchange networks in Sociology, Political Science & History. In contrast, the spread of the green squares in the attached figure comparing network authority with the 2010 NRC rankings appears to be broadly scattered, showing almost no connection between these two different data-based measures of departments.

While this network theory approach employs only one kind of data (the academic institutions that Ph.D. students later appear at as Ph.D. advisers), this direct data about the academic job prospects of degree holders should arguably be more important to prospective students than any of the 20 variables used in the 2010 NRC assessment to measure the “quality” of a graduate program.



Mathematical Genealogy & Department Prestige

Sean Myers[1], Peter J. Mucha[1], and Mason A. Porter[2]

[1] *University of North Carolina, Chapel Hill, NC 27599, USA*

[2] *University of Oxford, OX1 3LB, UK*

(Received ??/??/????; accepted ??/??/????; published online ??/??/????)

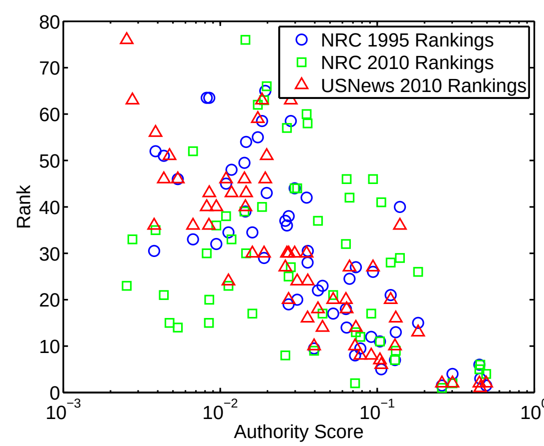
[DOI stuff]

The Mathematics Genealogy Project (`http://www.genealogy.ams.org/`) is a database of over 150,000 scholars with advanced degrees in mathematics and related fields. Entries include dissertation titles, adviser(s), graduation years, degree-granting institutions, and advisees. The MGP is popular among mathematicians, and it can be used to trace academic lineages through luminaries like Courant, Hilbert, and Wiener to historical predecessors such as Gauss, Euler, and even Kant. For example, MGP data was used recently to study the role of mentorship in protégé performance [1].

We consider recent branches of this mathematical family tree by projecting the MGP data for degrees granted since 1973 onto a network whose nodes represent academic institutions in the United States. An individual who earns a doctorate from institution A (during the selected period) and later advises students at institution B is represented by a directed edge of unit weight pointing from B to A. The total edge weight from B to A counts the number of such advisers.

This network representation can be used to estimate the mathematical prestige of each university using various "centrality" scores [2] of the corresponding node. We represent "hub" and "authority" scores [3] using node size and color (red to blue), respectively. Institutions with high authority scores have high-valued hubs pointing to them, and high-valued hub nodes point to high-valued authorities. In the legend above, we list the top 20 institutions in order of their authority scores. We use a "geographically-inspired" layout to balance node locations and node overlap. A Kamada-Kawai visualization [4] places the high-authority universities in the network's center.

We compare authority scores with three rankings of mathematics departments [5–7] for the 58 universities that appear in the top 40 of at least one of the rankings or have one of the top-40 authority scores (see the scatter plot). As expected, higher authority scores correlate with higher prestige (i.e., smaller rank numbers). However, scatter is obviously present, particularly with the 2010 National Research Council (NRC) rankings.



[1] R. D. Malmgrem *et al.*, *Nature* **465**, 622 (2010).
[2] M. E. J. Newman, Networks: An Introduction, Oxford University Press (2010).
[3] J. M. Kleinberg, *J. ACM* **46**, 604 (1999).
[4] T. Kamada and S. Kawai, *Inform. Process. Lett.* **31**, 7 (1988).
[5] National Research Council 1995, `http://www.ams.org/notices/199512/nrctables.pdf`.
[6] National Research Council 2010 (rank order of medians of the S-ranking ranges; original release), `http://graduate-school.phds.org/rankings/mathematics`.
[7] US News & World Report 2010, `http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-mathematics-programs/rankings`.
[8] We thank Mitch Keller at the MGP for providing data. This work was supported by the NSF (PJM: DMS-0645369) and the James S. McDonnell Foundation (MAP: #220020177).