University of Oklahoma

Honors Math Thesis

MATH 3980

Ranking Math Departments Using the Google PageRank Algorithm

Author:
Teresa Ratashak

Supervisor:
Dr. Jonathan Kujawa

February 7, 2014

Abstract

Currently, math graduate departments are ranked subjectively by department heads and directors of graduate studies, so we tried to rank math departments objectively using the Google PageRank algorithm. This algorithm ranks webpages based on the number and quality of the pages that link to a particular webpage. We explored how the Google PageRank algorithm can be applied to ranking math graduate departments by using Ph.D. graduates as the linking structure. Specifically, we used two distinct data sets of Ph.D. graduates for this ranking; one data set contains information on where Ph.D. graduates obtained positions immediately following graduation, and the second data set contains information on where tenure and tenure-track professors at the Public Large Group universities and the Private Large Group universities, as defined by the American Mathematical Society, obtained their Ph.D. degrees. We computed different rankings using the algorithm, but because of the differences between webpages and math departments, we were not able to come up with a ranking that adequately compared math programs of substantially different sizes.

1 Background

The Google PageRank Algorithm is what sets Google's ranking of webpages apart from other search engines. This algorithm determines rankings by using the hyperlink structure of the Internet. A page is given importance based on the number and importance of pages that link to it. The rationale behind this is that when a page links to another page, the first page is endorsing the page it links to; however, a link from the Economist website should have more weight than a link from a blog that has three followers, which is why it is necessary to account for the importance of each linking page also. Furthermore, a page gives an equal proportion of its importance to each page it links to.

Specifically, the importance of a page P_i can be calculated using the following formula [2]:

$$I(P_i) = \sum_{P_j \in A_i} \frac{I(P_j)}{\ell_j}$$

where A_i is the set of all pages that link to page P_i and ℓ_j is the number of outlinks on page P_j

2 The Mathematics Behind the Algorithm

Instead of computing each page's importance independently, we can use matrices.

Define Hyperlink matrix

$$H_{ij} = \begin{cases} \frac{1}{\ell_j} & \text{if there is a link from } P_i \text{ to } P_j \\ 0 & \text{otherwise} \end{cases}$$

Define Importance vector $I_i = I(P_i)$, the importance of page P_i

Because we do not know the values of the importance ranking at the beginning, we will start by making each entry in $\mathbf{I}$ equal to $\frac{1}{n}$, where n is the total number of pages; this gives each page equal importance at the beginning, as well as ensuring that the importances sum to 1.

Now we can represent pages' importances with $\mathbf{I} = \mathbf{HI}$, and we can compute this with iterations $\mathbf{I}^{k+1} = \mathbf{HI}^k$.

To accurately compute importance rankings, we must make a few adjustments to ensure $\mathbf{I}$ is calculated correctly and will converge.

2.1 Adjustments

Some websites are dangling nodes, which means they do not have any outlinks. As the iterations go on these websites will collect importance without giving any away. We do not want this rank sink to occur, so we make our first adjustment by acting as if a page with no outlinks actually links to every other page. From a webpage perspective, this change can be justified because if a user browses a page with no outlinks, then the next webpage they visit they select randomly.

Define This adjusted matrix will be called $\mathbf{S}$ where

$$S_{ij} = \begin{cases} H_{ij} & \text{if } P_j \text{ has at least one outlink} \\ \frac{1}{n} & \text{otherwise} \end{cases}$$

To make sure the importance vector will converge, we need to make a few adjustments to make the matrix $\mathbf{S}$ stochastic, irreducible, and aperiodic due to Markov chain theory [6]. To accomplish this, we need to change our matrix in the following way:

Define $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\frac{1}{n}\mathbf{1}$, where $\mathbf{1}$ is an nxn matrix where every entry is 1, and $0 < \alpha < 1$ [2].

A matrix is column stochastic if the elements in each column sum to one [4]. Note that by fixing the problem of dangling nodes, we have made $\mathbf{S}$ column stochastic. $\frac{1}{n}\mathbf{1}$ is also column stochastic, since each column sums to $\frac{1}{n} \times n = 1$. Therefore, the combination of these two matrices creates a matrix $\mathbf{G}$ that is column stochastic.

A square matrix is irreducible if its directed graph is strongly connected. Because all entries of $\mathbf{G}$ are positive, this shows that every page is connected to every other page, which proves $\mathbf{G}$ is strongly connected and thus irreducible [2].

A matrix is aperiodic if it self-loops. Since the entries on the diagonals are each greater than zero, this shows $\mathbf{G}$ self-loops, and is thus aperiodic [6].

With this change, the importance matrix is guaranteed to converge. In order to justify this change, Brin and Page describe a situation where a user becomes bored and decides to abandon the link structure of the web and go to a random website [3]. Note that the closer α is to 1, the more weight is being placed on the link structure of the web, and conversely, the closer α is to 0, the more weight is being placed on randomness. Also, the smaller α is, the faster the importance vector converges. In general, for ranking webpages, Brin and Page use $\alpha = .85$ in order to balance the needs of converging quickly and giving more weight to the link structure of the web [2].

3 Relation to Ranking Math Departments

Currently, math departments are ranked based on ratings assigned by math department heads and directors of graduate studies; these people rate programs based on how good they think each program is. However, we want to see if there is a more objective way to rank math departments by looking at where Ph.D. graduates obtained positions.

To apply this algorithm to ranking math departments, we will use Ph.D. graduates as the links between math graduate departments. This means math departments will gain importance based on where their graduates are hired for postdoctoral work or professorships, which is similar to webpages gaining importance based on which webpages have outlinks pointing to them.

Our previous adjustments to the PageRank Algorithm also make sense for ranking math departments; if a math department hires no one, then we act like they are voting evenly for everyone to take care of the problem of dangling nodes. Furthermore, the randomness factor can still be added because people decide to accept a position for various reasons, not necessarily solely on the ranking of the university, so there is some randomness involved in the hiring of math graduates.

4 Process of Ranking

Before we begin ranking, we need to make a few remarks. First, we are only looking at Ph.D. granting math graduate departments in the United States, excluding applied mathematics programs. Secondly, we are ranking math departments based on how they link back into math departments, so we are excluding graduates who go into industry. This means the ranking will show which math departments give the best preparation for a position in academia, not necessarily which math departments are the best overall.

We are looking at two distinct data sets for these rankings. We used php and mysql to organize the data into a matrix representing the linking structure between the math departments for each data set. After this, we modified Jeremy Kun's open source Mathematica code and used it to run the PageRank Algorithm on these matrices [5].

4.1 American Mathematical Society Data Set

The first data set was provided by the American Mathematical Society (AMS), and it contains job placement directly following graduation, as it was reported to the AMS. This data set contains 2419 data points from 185 universities from years 2001 to 2011. In this data set, a link counts as obtaining any position, whether a postdoctoral fellowship, lectureship, or professorship, at a math department in the original set.

Define AMS Linking Matrix $L_{ij} = \Big\{$number of graduates school j hired that graduated from school i

Our results from the first time we ran the PageRank Algorithm on the data from the AMS are shown in Table 1.

Table 1: Initial Results for AMS Data with $\alpha = 0.9$

Ranking	**Name of University**
1	Washington State University
2	UC Berkeley
3	Missouri University of Science and Technology
4	MIT
5	University of Michigan, Ann Arbor
6	UCLA
7	Harvard University
8	University of Chicago
9	Princeton University
10	University of Texas, Austin
11	University of Maryland, College Park
12	Columbia University
13	Cornell University
14	University of Colorado, Boulder
15	New York University, Courant Institute
122	University of Oklahoma

Looking closely at this initial ranking, we realized Washington State University, the university that is ranked number one, only has six votes, four of which are from itself. Therefore, using this ranking system, that university was able to inflate its own ranking. Hopefully every university believes that their graduates are well-prepared and would thus hire their own graduates, so we decided to take out all self-votes from the matrix so that a university cannot affect its own ranking in that way. Below are the new results for $\alpha = 0.90$ and $\alpha = 0.85$.

Table 2: AMS No-Self-Voting Results $\alpha = 0.9$

Ranking	**Name of University**
1	UC Berkeley
2	MIT
3	University of Michigan, Ann Arbor
4	University of Chicago
5	UCLA
6	Harvard University
7	Princeton University
8	Columbia University
9	University of Texas, Austin
10	University of Illinois, Urbana-Champaign
11	Cornell University
12	University of Wisconsin, Madison
13	University of Maryland, College Park
14	Purdue University
15	New York University, Courant Institute
95	University of Oklahoma

Table 3: AMS No-Self-Voting Results $\alpha = 0.85$

Ranking	**Name of University**
1	UC Berkeley
2	MIT
3	University of Michigan, Ann Arbor
4	University of Chicago
5	UCLA
6	Harvard University
7	Princeton University
8	University of Texas, Austin
9	Columbia University
10	University of Illinois, Urbana-Champaign
11	University of Wisconsin, Madison
12	Cornell University
13	Purdue University
14	University of Maryland, College Park
15	New York University, Courant Institute
92	University of Oklahoma

Comparing Table 2 with Table 3, there are few differences in the ranking, even though Table 2 uses $\alpha = 0.9$ and Table 3 uses $\alpha = 0.85$. In fact, none of the top 15 universities moved more than one place in ranking. For the rest of this paper we will use $\alpha = 0.9$ because the ranking does not change much from these two values of α and because $\alpha = 0.9$ places more weight on the link structure between math departments than $\alpha = 0.85$.

After looking at these new rankings, we noticed that many of the bigger universities were ranked in the top 15. Looking closer, we discovered that all 13 universities with more than fifty graduates linking into the matrix were ranked in the top 15 schools. This makes sense because schools are given a vote for each graduate of theirs that is hired by another university in the matrix, the weight of the vote depends on the importance of the hiring university, but more graduates will still generally add up to gathering more importance overall. This is another way in which this situation is different from webpages because a webpage has the potential for any number of links pointing toward it, but universities graduate different numbers of students depending on size and funding. Therefore, if we left the ranking this way, there is no way to compare universities of different sizes.

To account for this, we decided to try taking size completely out of the equation, by allowing every university to receive exactly one vote, which was weighted according to the different universities that hired these graduates. To do this, we made $\mathbf{L}$ row stochastic; each row shows the votes for a particular school, so we divided each row in the matrix by the number of graduates. The results from running the PageRank algorithm are in Table 4.

However, now these results do not give universities any advantage for having a bigger graduating class. Having a bigger graduating class can be a reflection of a successful university that is attracting and retaining more students, so this should not be completely disregarded. At this point, though, there is no way to objectively find this balance. However, we could look solely at the Public Large

Table 4: AMS Weighted by Size $\alpha = 0.90$

Ranking	**Name of University**
1	University of Florida
2	Stony Brook University
3	Northwestern University
4	University of South Carolina
5	Texas A & M University
6	UC San Diego
7	UC Irvine
8	University of North Carolina, Chapel Hill
9	University of Nebraska, Lincoln
10	Arizona State University
11	University of Southern California
12	Purdue University
13	University of Iowa
14	Tulane University
15	Tufts University
71	University of Oklahoma
84	Harvard University
85	MIT

Group and the Private Large Group universities, since these universities are more comparable in size. These are the 26 public university math departments and 24 private university math departments which had the highest average number of Ph.D.'s awarded between 2000 and 2010 [1]. The results from running the PageRank algorithm on this data set containing 1250 data points are in Table 5.

Table 5: AMS Public and Private Large Groups Results $\alpha = 0.9$

Ranking	**Name of University**
1	UC Berkeley
2	MIT
3	Harvard University
4	University of Michigan
5	UCLA
6	Princeton University
7	University of Chicago
8	University of Texas, Austin
9	Columbia University
10	New York University, Courant Institute
11	University of Maryland, College Park
12	Brown University
13	Cornell University
14	University of Wisconsin, Madison
15	University of Illinois, Urbana-Champaign

4.2 Tenure-track Professor Data Set

I obtained the second data set by gathering data online in Fall 2012, and it contains the alma mater of tenured and tenure-track faculty at the Public Large Group and the Private Large Group universities. This data set contains 1398 data points from 50 universities, with professors earning their doctoral degrees anytime between 1948 and 2011.

For the second data set, after taking out self-voting and only keeping the data points which link back into the matrix, which are those tenure-track professors who graduated from Public or Private Large Group universities, the ranking results are shown in Table 6.

Table 6: Professor No-Self-Voting Results $\alpha = 0.90$

Ranking	**Name of University**
1	Princeton University
2	Harvard University
3	UC Berkeley
4	MIT
5	Stanford University
6	New York University
7	University of Chicago
8	UCLA
9	Columbia University
10	California Institute of Technology
11	Yale University
12	Brandeis University
13	University of Washington
14	Rutgers University, New Brunswick
15	Cornell University

For this data set, we are only looking at the Public Large Group and the Private Large Group universities. Since these two groups are made of the largest universities, size differences do not skew the data. However, this data set has the unique characteristic that the tenure and tenure-track professors graduated between 1948 and 2011; when a professor earned his or her degree should be taken into account because professors who recently earned their degrees should more accurately reflect the quality of current graduate programs than professors who earned their degrees earlier. To account for this, we weighted the votes by decade, as described in the chart below, and Table 7 shows the ranking after this modification.

Years	2000 & later	1990-1999	1980-1989	1970-1979	1969 & earlier
Weight of Vote	1	.8	.6	.4	.2

Table 7: Professor Weighted by Decade Results $\alpha = 0.90$

Ranking	**Name of University**
1	Princeton University
2	Harvard University
3	UC Berkeley
4	MIT
5	Stanford University
6	University of Chicago
7	Columbia University
8	California Institute of Technology
9	New York University
10	UCLA
11	Yale University
12	University of Washington
13	University of Pennsylvania
14	Rutgers University, New Brunswick
15	Cornell University

5 Interpretation of Results

Now we need to compare the differences between the weighted and no-self-voting versions of each data set.

For the data about positions directly following graduation, the no-self-voting results in Table 2 have only one university, Purdue University, in common with the weighted-by-size results in Table 4. One possible supposition from these results is that larger schools do not prepare their students as well as smaller universities for jobs after graduation. The graduates from smaller universities may connect to more highly ranked universities because students may receive more personal attention at a smaller school and thus be better prepared for academia afterward.

However, the data we are looking at only contains information for graduates from Ph.D. granting American universities who are employed at another Ph.D. granting American university. This means that this data does not contain employment information for all graduates; in fact, 10 out of the top 15 universities from the weighted-by-size results had fewer than 20 percent of their graduates obtain positions at linking universities, while all of the top 15 universities from the no-self-voting results had at least 20 percent of their graduates obtain positions at linking universities.

Therefore, it is possible that the graduates who linked back into the matrix and affected the ranking were the top students for the top universities in the weighted-by-size results, while the graduates who linked back into the matrix and affected the ranking were more of an average graduate for the top universities in the no-self-voting results. In the weighted-by-size ranking when every university only received one total vote rather than one vote per graduate, the comparison may really be between the top students at one university and the average graduate at another university. Therefore, this comparison may not be an accurate reflection of the relative standing of universities either.

It seems that a more accurate ranking using the AMS data would take size

into consideration, but would not allow sheer size to overwhelm the ranking. We could not think of an objective way to do this for all universities, but we did look at the subset of the AMS data which only includes graduates from Public and Private Large Group universities that link back into Public and Private Large Group universities. The only universities that are different between the no-self-voting top 15 ranking and the large-group-only top 15 ranking are Brown University and Purdue University; the rest of the universities are in both top 15 rankings, though they may have moved up or down in ranking. Even though Brown University did not appear in the top 15 ranking for the no-self-voting results, Brown University was ranked number 16 in those results, so going from 16 to 12 is not a huge change. However, Purdue University went from being number 14 in the no-self-voting ranking to number 23 in the large-group-only ranking. Looking closer, out of the 58 Purdue graduates in the AMS data set, only 26 graduates linked to a Private or Public Large Group university. Therefore, this drop in Purdue University's ranking can be explained by the fact that over half of their graduates in the AMS data set linked to universities that did not affect the large-group-only ranking. This is a reminder that the large-group-only ranking represents which universities in the Private Large Group and Public Large Group best prepare their students for positions at other Private Large Group and Public Large Group universities, not necessarily which universities best prepare their students for academia in general.

However, we were unable to come up with a robust ranking for the complete AMS data set because we were unable to adequately compare math departments of substantially different sizes.

For the second data set, assigning values for different decades was subjective, so it is important to note that the ranking weighted by decade is not completely objective either. However, the top five universities for both rankings of the second data set are the same, and the only universities that are different in the top 15 are the University of Pennsylvania and Brandeis University. The University of Pennsylvania went from being number 18 to number 13 in the weighted ranking, and out of the 20 graduates from this university, 9 graduated in the 2000s, so this is why the weighting increased the University of Pennsylvania's ranking. Also, Brandeis University went from being ranked 12 to being ranked 19; the most recent graduate from Brandeis University included in the rankings was from 1996, so this helps explain why this university decreased in ranking after accounting for when people earned their doctoral degrees. Because the two major differences between the rankings logically make sense, the weighted-by-decade ranking seems to be a good ranking of the universities.

Since we have compared the results within each data set, now it is time to compare the results between the two data sets. The AMS large-group-only results and the tenure-track professor weighted-by-decade results are ranking the same 50 universities and thus should be comparable. However, only nine of the universities between the two top 15 lists are the same, and some of the ranking differences are much different, such as Stanford University being ranked fifth in the professor results and twenty-fifth in the AMS results. I believe a big factor in the discrepancy between the two rankings is the nature of first placements after graduation versus tenure-track positions. The AMS data on the Public and Private Large Groups contains 1250 data points from a span of 11 years, while the Tenure-track Professor data on these universities contains 1398 data points from a span of 64 years. This shows that there is a limited number of

tenure-track positions available; not every graduate who attains a postdoctoral fellowship, professorship, or lectureship right after graduation will end up becoming a tenured professor at a Public or Private Large Group university. It is possible that the professor ranking more accurately represents the quality of graduate programs because the limited number of tenure-track positions might make quality outweigh quantity of graduates. For example, Stanford University only had 21 graduates link into the large-group-only AMS data set, but they had 69 tenure-track professors at linking universities in the professor data set; although the number of graduates in the last decade is relatively small compared to schools in the top 15 AMS results, the university has a remarkable number of graduates in tenure-track positions which implies that a high percentage of their graduates go on to attain tenure-track positions.

On the other hand, it can be argued the AMS large-group-only results may more accurately portray the current quality of a graduate math department since the data is from 2001 to 2011, while the data for the tenure-track professor results are from 1948 to 2011. This is especially true for math departments that are drastically changing in quality. If a graduate math department starts to drastically improve or worsen, then the AMS data would capture this change much more quickly than the tenure-track professor data, since it takes time for Ph.D. graduates to attain tenure-track positions, but graduates can immediately attain professorship, fellowships, and lectureships. This means another explanation for the discrepancies between the two top 15 rankings is that the universities in the AMS top 15 that are not in the tenure-track top 15 may be up-and-coming universities, while universities that appear in the tenure-track top 15 and not in the AMS top 15 may be declining in quality in recent years.

These two theories offer very different interpretations of the rankings from the different data sets. However, the answer may not simply be just one or the other, but may be a case-by-case look at the universities that appear highly in only one of the rankings.

6 Conclusion

Both of these data sets differ in fundamental ways from webpages, which is why the PageRank algorithm has not been able to give us a perfect ranking of the math graduate departments. First of all, a webpage can have an outlink to a second webpage without the second webpage doing anything, and because of this a webpage can always vote for whichever webpages that it thinks are the best. However, a math department can only hire, and thus vote, for a graduate who first decides to apply for and accept a position. People decide to accept positions for multiple reasons, such as geographical location and family, which do not relate to the prestige of a university, so ranking is not as simple as with webpages.

Furthermore, a Ph.D. program is not the only thing that shapes a person; undergraduate work and experience also affect employability, and there is not a good way to account for these other factors in the ranking. While there are also additional factors to consider in ranking webpages for search results, such as the font size of the search words, these factors are incorporated by Google, but not by the PageRank algorithm itself [3]. Also, font size is much more easily measured objectively than each individual's undergraduate experiences.

Also, as we have already mentioned, different universities have different sizes

of programs and are thus limited in the number of students they can graduate in a given year; this affects how many votes a university can receive. However, there is not a limit on the number of webpages that can link to a specific webpage, which is why this difference is also hard to account for when using the PageRank Algorithm. Therefore, because of these complicated differences between the linking structure between math departments and webpages, we have not been able to give a robust ranking of math departments, especially math departments of different sizes.

7 Acknowledgments

I would like to thank Jim Maxwell from the American Mathematical Society for providing much of the data that I used for this project. I would also like to thank Ralf Schmidt from the math department for his patient guidance with using mysql and php. Finally, I would like to thank Jonathan Kujawa for the insight and support he has given me as my honors research adviser.

References

[1] American Mathematical Society, cited 2013: *Annual Survey Groupings of Doctoral Departments*. Available online at http://www.ams.org/profession/data/annual-survey/groups.

[2] D. Austin, cited 2013: *How Google Finds Your Needle in the Web's Haystack*. Available online at http://www.ams.org/samplings/feature-column/fcarc-pagerank.

[3] S. Brin and L. Page, cited 2013: *The Anatomy of a Large-Scale Hypertextual Web Search Engine*. Available online at http://infolab.stanford.edu∼backrub/google.html.

[4] K. Bryan and T. Leise, cited 2013: *The $25,000,000,000 Eigenvector: The Linear Algebra Behind Google*. Available online at http://www.rose-hulman.edu/∼bryan/googleFinalVersionFixed.pdf.

[5] J. Kun, cited 2013: *Google's PageRank- The Final Product*. Available online at http://jeremykun.com/2011/06/20/googles-page-rank-the-final-product/.

[6] A. N. Langville and C. D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, Princeton, NJ, 2006.

[7] A. N. Langville and C. D. Meyer, *Who's #1? The Science of Rating and Ranking*, Princeton University Press, Princeton, NJ, 2012.