Citation and Affiliation Networks in Academia

Gabriel Hobeika
ECE, Carnegie Mellon University
Pittsburgh, PA, USA
gph@andrew.cmu.edu

Dillon Lareau
ECE, Carnegie Mellon University
Pittsburgh, PA, USA
jlareau@andrew.cmu.edu

Abstract

Citation networks in academia can provide insight into the success of papers, as well as the success of their authors. Furthermore, trends in paper citations can point to new rising ideas, enlighten our under-standing of relationships between universities and other institutions and between disciplines. While there has been some research into academic citation net-works, we feel the need to go deeper. We will browse databases of papers, scrape relevant information to create networks, and then analyze those networks to identify what creates a successful author, as well as any interesting relationships between institutions, journals and cross discipline works. We will compare our work and the network data that we find against other metrics that measure success. We will then see if any network-science based metrics help us predict or identify success in the traditional networks.

1. Introduction

Our project will be looking at network trends in the world of academia. More specifically, we seek to ana-lyze citations and references between different au-thors, disciplines, and fields of study. We hope to be able to identify what marks success in different ways throughout academia as described by our questions below.

2. Motivation

Firstly, we specifically want to analyze what types of colleges and universities are most likely to have au-thors cross collaborate. We also hope to identify whether things like ranking, endowment, or school type (private vs. public) determine anything about whether academics are likely to work together in the future. We also want to see how ranking of university is related to institutions that they work with, and see if we can devise a way to rank those institutions based on their network position.

Secondly, we want to see what journals are most likely to cite one another. We are specifically interested in whether the prestige of a specific journal affects the ratio of self-citing versus inter-journal citing. We also want to see what, if any implications, our network can show for conferences, and one-off publications.

Thirdly, we want to identify how network statistics relate to author success. More specifically, we would like to evaluate how citations networks, and collabora-tion networks affect metrics like H-index. This will give us an idea of what authors should do in order to increase their H-index in the future and should help us identify when an author is poised to be more success-ful and get a higher H-index.

Lastly, we want to be able to look at some of our pre-vious subjects across time. For instance, we would like to see how an authors position in the network at the time of publishing a paper relates to their success in future publications contained in the network that we created.

Networks are fundamental to each of these goals. For each of the goals, in order to identify the relationships between papers on a large scale we will have to create networks that identify what papers cited, where those papers are from, whom they reference, and what con-nections do the papers and authors have to the outside world.

3. Prior Work

We have viewed some authors previous work in the field of academic paper citation networks.

One of our primary inspirations titled A Century of Physics [1], looks at 100 years of papers in the physics world through web of science and analyzes the long-term impact of different publications based on their reference data. It shows how publications and papers can have an impact over a range of time.

Our work differs from that paper, as we are generally looking at small-scale in terms of time when compared to the 100 years of physics papers. We want to be able to identify what will cause an impact, not what has caused an impact in the past. Additionally, as the title implies, A Century of

Physics focuses just on the physics discipline. It is our hope to find data that can apply to multiple disci-plines, and cross discipline work.
We also looked at Newmans [2] approach in Finding community structure in networks using the eigenvectors of matrices. This paper was helpful in helping us analyze the citation information from a community perspective. Additionally, Newman has a publicly available dataset for his work, which we do not directly use in our work, but that we have viewed to see how networks are sometimes structures when applied to academic citations and collab-orations.
We differ from Newman, as we are not just concerned with communities in academia, but a concept of success. As such, we are interested in more detailed information about the papers themselves, and what topics they cover, not just the citations and cross-references between differ-ent academics.
We want to focus on broad analysis of the networks in academia, over the whole spectrum of academia. Most other work in the field focuses on one metric, or one quantitative finding. Our work seeks to define the abstract relationships across academia, between journals, institu-tions, and even between authors themselves. We hope that our broad scope will give us insights into success not pre-viously explored.

4. Approach

Our approach relies on our attempts to acquire a wide range of data. As highlighted before, we need many data to evaluate the metrics we want, and as such, we are going to need to figure out ways to extract data from exist-ing databases for usages that they have not catered to be-fore.
As such, during milestone one, and future milestones we developed ways to scrape data, and use APIs for different databases to grab the information that we need. We faced limitations in our data collection during the first mile-stone, that in the second milestone we have worked around. We also have been looking at previous works to identify what exactly we need that is different from what they have provided.
The databases that we used were the SCOPUS database, and the SAO/NASA Astrophysics system. Scopus allowed us to get data for overall institution collaborations. SAO/NASA made it convenient to get individual paper data. Our methodology for the SAO/NASA Data Collection was as follows: First start with some subset of 10 papers, then go on to get all the papers those cite, and finally get the second level of citations for those next papers. For Scopus we started with an initial institution. We then found the top 150 collaborating institutions of that institution. Then gathered the top 150 collaborating institutions of those institutions.
We used the Python NetworkX package to construct our networks from the raw data; the construction of each individual network will be elaborated on further in the Experimental Setup sections. We also used Gephi to visualize our networks and aid in community detection.
After we finished preliminary data collection, we began to provide rudimentary network analysis on the net-works we constructed. This involved looking at average degree and distribution of nodes. It also involved using gephi to identify the community groups within the different networks we had developed.
After completing our rudimentary analysis we began looking at more in depth metrics to analyze our networks. Following the rudimentary analysis, we will then begin our in depth analysis which will involve creating heuris-tics to identify what makes a successful paper. We will also begin diversifying our networks. For instance, one network may have nodes of authors with weighted edges of coauthorships between them, where as another one may have a paper, and all the citations of that paper, while still another may have an institution with weighted edges of coauthorships to another institution. We will create an initial set of these diverse networks for milestone 2.
For milestone 3 we have done further analysis on the networks we created for milestone 2. We have now taken an in depth look at the community structure of the networks and picked specific nodes to examine so that we can further understand what our networks tell us about academic citations.
After doing all this data collection and network creation, we hope to be able to analyze the data and truly identify the trends in academia that we were looking for. We are assuming that many of these networks will have some very connected nodes and some fringe nodes, as we be-lieve that academic work comes in bursts, and some pa-pers are fundamental to academia. We have also made the assumption that cross discipline work is prevalent in aca-demia and will be easy to identify. We also have gone into this project with the assumption that private and pub-lic universities may have different types of citation net-works. We are no longer assuming a U.S. centric data set will be useful in finding our results. This is after analyz-ing the networks we currently have; and seeing that aca-demic work is far more internationally connected than we originally thought.

5. Results

5.1. Experimental Setup and Practical Results M1

Through our initial research and data gathering, we have found that most papers include a large number of cita-tions, between 10 and 20 citations a paper. We gathered our initial set of papers from the SAO/NASA Astrophys-ics Data System [3]. Because the number of citations was so large,

we had to scale down our network in order to run some simulations on it on GePhi. What we found in our citation network was rather interesting. The network end-ed up having 6,808 nodes and 10,607 edges. This graph had an average degree of 3.116 and the distribution of the degrees led us to believe the graph was scale free. The graph also had an average clustering coefficient that was lower than we were expecting at 0.047. Lastly we found that upon using a random algorithm to induce communi-ties, we had around 19-21 communities form. Essentially, there was a lot of cross connection between papers in our network. We also did see what we are calling fringe papers. Those papers are papers on the edge of the net-work that cite a large number of papers not found in the rest of the network. A picture of this network is included in the figure below.

5.2. Experimental Setup and Practical Results M2 + M3

After getting that initial research from M1, we decided to narrow down what we were looking for, and we split up our data collection so that we could accomplish all of our goals. What we looked to create were:
1. A collaboration network between authors who have at least 10 publications
2. A citation network between journals based on the data collected regarding author citations.
3. A collaboration network between institutions that have at least 50 papers each.
We also kept our initial paper citation network as a reference point, but we did not use it for any meaningful re-sults during these milestones
We managed to obtain more data out of the SAO/NASA Data system. Using the data that we obtained, we created both a collaboration network between authors who have at least 10 publications, and a citation network between journals based on the data that lead to the author network. A note: We had to limit the author collaboration network by not including two specific papers on the Higgs-Boson. In milestone 3 we have attempted to rectify this by weighting the edges in that network so that high-author papers have less controlling influence on the network structure as a whole.
We chose not to include those papers, as the amount of authorships in those papers was skewing the network. For reference, both papers needed appendixes to note all of their authors, of which they had more than 2000. We do believe these two papers are still relevant to our research, as they do show that significant papers in academia can sometimes be the result of concerted, many yearlong ef-forts. Those papers, because they are so fundamental to their field, will usually be well connected, and heavily cited within academia. Additionally, these papers are unique to our dataset, as CERN, which funded them, is a multinational project, and the papers created on the Higgs-Boson from there must cite all researchers invited to spend time at CERN.
Firstly, we looked at the author collaboration network (pictured below). Our first approach was to look at the communities within this network. We found a high level of modularity, 0.831, and a high number of communities, with the normal random community-finding algorithm used in GePhi.

We also found that the average degree was about 27, with a network diameter of 11. We expected the average de-gree to be high, as we had limited this network to authors with over 10 papers, as such they should have a large number of collaborators.
The high modularity was also interesting to note. Some of the papers defined our communities. For instance, some papers had 40+ authors and those authors would all tend to form one community. When an author happened to work on two or more of those papers, they tended to be-come a central node, which defined a path between two communities.
We found GePhis community analysis helpful in finding these specific authors. We have looked at this as one met-ric of success within our network; that is to say if an au-thor collaborates on more than one big papers, he must be a respected author.
Next, we looked at a journal citation network (pictured below). It was generated from the set of papers that made the author collaboration network. Unlike the author net-work, this network did not have a high level of modulari-ty. In addition, it had a low level of communities, regard-less of the resolution parameter.

The modularity was a low 0.320. The average degree in this network was much lower than author citations, at 5.259, though the weighted degree was about where we expected it, at about 26.9. Of note, there was a low level of communities, as community structure did not particu-larly influence these journal citations.
The interesting result we found was when we compared the network we generated to the H5-Index found on google scholar.[7] We found our most heavily cited jour-nals, like the Astrophysics Journal, and the Monthly No-tices of the Royal Astronomical Society also scored high on the H-index rating. We think our measure of success can be slightly different from H5-Index, where H-Index looks at citations over time; our rating looks at bursts of citations in one particular field. For this particular field, astrophysics, our metrics did not completely align with H-Index. In particular, The Astrophysical Journal scored higher on our metrics than Science or Nature, which both have much

higher H-Indexes. We believe looking at burst citations may be valuable in establishing better success metrics for our papers.

In milestone 3, we refined these networks somewhat. We looked at physics collaborations vs astronomy collaborations during milestone 3. What we found was somewhat interesting. We started with the same amount of papers, and roughly the same amount of authors, the numbers were as follows:

TODO: insert chart

Using this data, we observed some interesting differences between the astronomy and physics datasets. The first and most obvious difference is the number of authors with ¿10 publications. Significantly more authors in astronomy seemed to have worked on that many publications within this collaboration network.

This was interesting, as we had previously figured that the structure of the citation networks would not vary that greatly between disciplines. This set of data, however, shows something different. It suggests that with physics papers there are less prolific authors among the data that we were able to sample, than astronomy datasets.

The physics dataset, filtered on 10 papers, also is not a completely connected network. The image below shows the comparative sparsity of the dataset when compared to the above images of the astronomy data. The image shows that in physics, like astronomy, there are clearly defined communities. That is to say that we identified that some authors are still more likely to work together and to cite each other and that within physics there are still some prolific authors.

We have also looked at H-Index as a success metric [6], both globally and measured in our own citation network. We compared this universal success metric to network factors, like node degree, K-Core, and centralities. We have included graphs with this paper to show the relevant trends that we have found. In summary, those trends are generally inconclusive. We were however able to find some interesting results when looking at collaboration within our network when compared to the H-Index of the network. These results will be discussed more after the charts.

As mentioned before, our most significant result when looking at H-index relative to the paper networks, was H-index to degree of collaborations. This graph, shown be-low, has a high degree of certainty ($R^2 = 0.8491$)

This graph shows H-index on the x-axis with degree on the y-axis. This is the most highly correlated graph we have found. What we can identify from this is that when an author collaborates heavily, and works with a diverse array of authors, they will be more likely to score high on the universal metric of h-index.

We believe this has some rational backings: first of all, if an author collaborates more, they may have published more papers. Publishing papers has a factor in h-index as it measures productivity of paper writers. Also, if a pro-fessors work is collaborated on by other professors, those professors may self-cite their work in other papers. This will lead to the second requirement of H-index, citations, to be accomplished more easily.

The last network we generated drew from a different source. We used SCOPUS [4], which is one of the largest citation databases, in order to find the top 150 collaborating institutions, from 50 starting institutions. We came to this method of data collection, because we found that in-stitutions typically have collaborations and citations with a very large number of other institutions. To try to grab all of the collaborations would be near impossible. We fig-ured the top 150, which typically meant more than 500 collaborations had occurred, would narrow down our data set to a manageable point, without removing too much specificity.

Still, when making this network (pictured below), we found that we had almost 700 very inter connected institu-tion nodes. As with the Journal Citation network, this collaboration network had a low level of modularity, and not a large number of communities. In fact, establishing a real community structure was almost impossible. We made sure to add weight to all of the edges, and when visualizing, made the most heavily connected nodes larg-er.

Because of the method of data collection used, we also decided to filter out those nodes outside of the top 100 most connected nodes (pictured below), most connected nodes for better visualization. We found this showed the communities in a slightly nicer way, but the communities were still not heavily modular and did not inform us of much.

What we did find, upon manual analysis of the data was relatively interesting. We found that despite looking at the top 150 collaborated institutions for a multitude of institu-tions, some remained at the top of the list across the board. Of note, the Ohio State University and Massachu-setts Institute of Technology both appeared at the top, or near the top of many of the institutions in our list. We also found that institutions heavily collaborated with institu-tions that were highly ranked in similar fields with them. For instance, Carnegie Mellons papers were 34.4% com-puter science papers, and their second most collaborated institution was MIT.

We also found that institutions tend to cite other institutions that are closer to them geographically more heavily, as one would expect. For example, CMU and University of Pittsburgh collaborated more than even CMU and MIT. Likewise, MIT and Harvard, which are both in Cam-bridge, Massachusetts worked together quite heavily.

The core metric, however, when looking at institutions

continues to be that institutions number of authors. Institutions with a large number of authors tend to just have more coauthorships in general. We somewhat accounted for this problem in the data collection method. Because our data collection was just based on the top 50 collabo-rating institutions for the institutions that we looked at, we were able to ensure that one institution with a very high number of publications wouldnt be significantly biased over the others. This is because we are not looking at all their papers, but rather at the subset of papers that were worked on with their most collaborated affiliations.

When comparing the data across the communities, and to publically available ranking data we did garner some in-teresting insights. Firstly, as mentioned before, institutions with high rankings in a particular field did tend to work together more often. We looked into this further, to see if this was biased by there being a large number of papers published by those institutions in those fields, and we found that wasnt necessarily the case. For instance, while roughly 34% of CMUs papers are published in the field of computer science, only about 10% of the papers from MIT were published in that field. That hints at the notion that CMU researchers may be more inclined to work with MIT researchers, and vice versa, even though CMU publishes more computer science papers compared to its other disciplines than MIT does.

During milestone 3, we also looked at our institution cita-tion data in conjunction with the ARWU university world rankings [5]. Of note, ARWU rankings claim to include publication citation data within their rankings of universi-ties. As such, we looked at a subset of universities from their rankings, and compared it to the larger nodes within our own network. Our results were somewhat compelling.

When looking at ARWU rankings, we saw that high ranked institutions, namely Harvard, and MIT appeared to be collaborated with by most of our institutions. That is to say, that these very highly ranked institutions merited more collaboration across the board. We also found hits of a rich-get-richer property of these networks, so good schools tended to like to work with other good schools. In fact, a key metric is that none of the universities who ap-peared in our top-50 list fell out of the top 500 universi-ties according to the ARWU.

In a word, these are interesting findings because it con-firms what we know about the schools; they are good schools and as such professors want to work with them more often.

It would seem now that the prestige of the institution is a factor in authors working with one another, we are not sure of the reason for this yet. We may hypothesize that high ranked institution relationships give some credence to the findings in the paper; for example, professors are more likely to give weight to a paper with an author from Harvard.

From these results, we can find one simple truth; when compared against general world University Rankings, one can be sure that working with a well-connected university will increase the amount of academics exposed to a paper, and will have the opportunity to work with more academics.

For the next milestone, we are looking at applying a page-rank like analysis to the rankings. We would like to see if position in the graph relative to high-ranking universities indicates anything about the ranking of your university. For instance, if one authors papers with primarily Har-vard, does that mean that their university is also good? This will serve as a way we can predict what a university may be ranked in the future. We also want to look at other datasets in order to confirm our results like the Times international rankings.

Our notions of success for institutions still need some polish, but we have found that highly ranked institutions, via world rankings (from Wikipedia) tend to cite other institutions that were also highly ranked. We also found that institutions with a large number of authors found themselves at the top of the collaboration lists (like the Ohio State University.) While not explored yet, we think that filtering on paper type and institutional rankings from third party sources may give us more incite in to what makes a successful institution successful.

6. Conclusion

We have found now, that because we were able to gather a significant amount of data that we are beginning to see some results. Those results do somewhat differ from the results we had anticipated. For instance, the institution citation network had to include international institutions rather than just US based ones. Without internationally institutions, it would have been impossible accurately capture the collaborations that occur between institutions. International collaboration is simply much more common than we thought.

We did however see some results that we did expect. For instance, having a strong community structure between authors, when based on collaborations is something we did expect. That, as mentioned, derives itself from the fact that some authors worked on multiple highly popular and highly collaborated on papers. These authors became central nodes in the collaboration graphs.

We still feel that our metrics for success can be further refined. We must find better metrics than strong commu-nities must, and high degrees to determine which of our nodes are truly successful, though we believe this is a good starting point.

Interestingly for our university data, we found what one

would expect, that high ranked universities are likely to be collaborated with across the board. This could lead us in to our first unique success metric; that working with well-regarded universities as viewed by typical rankings tends to make an author more well connected.
Our next step in that regard should be to classify the universities more tightly, and maybe establish a ranking system that shows how many times they are cited by other high ranking universities. This may help us identify insti-tutions are up and coming, and ones that may make a splash in college rankings. We will also produce these same results with H-index

7. Future Work

M3: Now that we have collected a good amount of working data, we must begin thinking more in depth how to compare success in academia. We should also look into cross correlating both H-index, and magazine rankings with our journal and institution networks. Lastly, we will work together on branching out the author subjects from just the field of astrophysics.
M4: Further analysis of our data from M2 left us with some interesting questions. We are now narrowing down the success metrics we might be using to view institution-al and journal success. We need to apply these similar metrics to authors. We hope to be able to show by the deadline some unique metrics that will truly help us iden-tify what success is in academia, and how to see when success will be achieved. For instance, the colocation of the universities with good universities may be im-portant, we also plan to do this analysis with h-index of authors.

8. Citations

[1]Sinatra, Roberta, et al. "A century of physics." Nature Physics11.10(2015):791-796.

M. E. J. Newman, Phys. Rev. E 74, 036104 (2006

http://www.adsabs.harvard.edu/

http://www.scopus.com/

http://www.shanghairanking.com/ARWU-Statistics-2016.html

https://en.wikipedia.org/wiki/H-index

https://scholar.google.com/citations?view$_op$ = $top_venueshl = en$

8.1. Type-style and fonts

Figure and table captions should be 10-point Helvetica boldface type as in

Figure 1. Example of caption.

Long captions should be set as in

Figure 2. Example of long caption requiring more than one line. It is not typed centered but aligned on both sides and indented with an additional margin on both sides of 1 pica.

Callouts should be 9-point Helvetica, non-boldface type. Initially capitalize only the first word of section titles and first-, second-, and third-order headings.

8.2. Footnotes

Please use footnotes sparingly[1] and place them at the bottom of the column on the page on which they are referenced. Use Times 8-point type, single-spaced.

[1]Or, better still, try to avoid footnotes altogether. To help your readers, avoid using footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence).