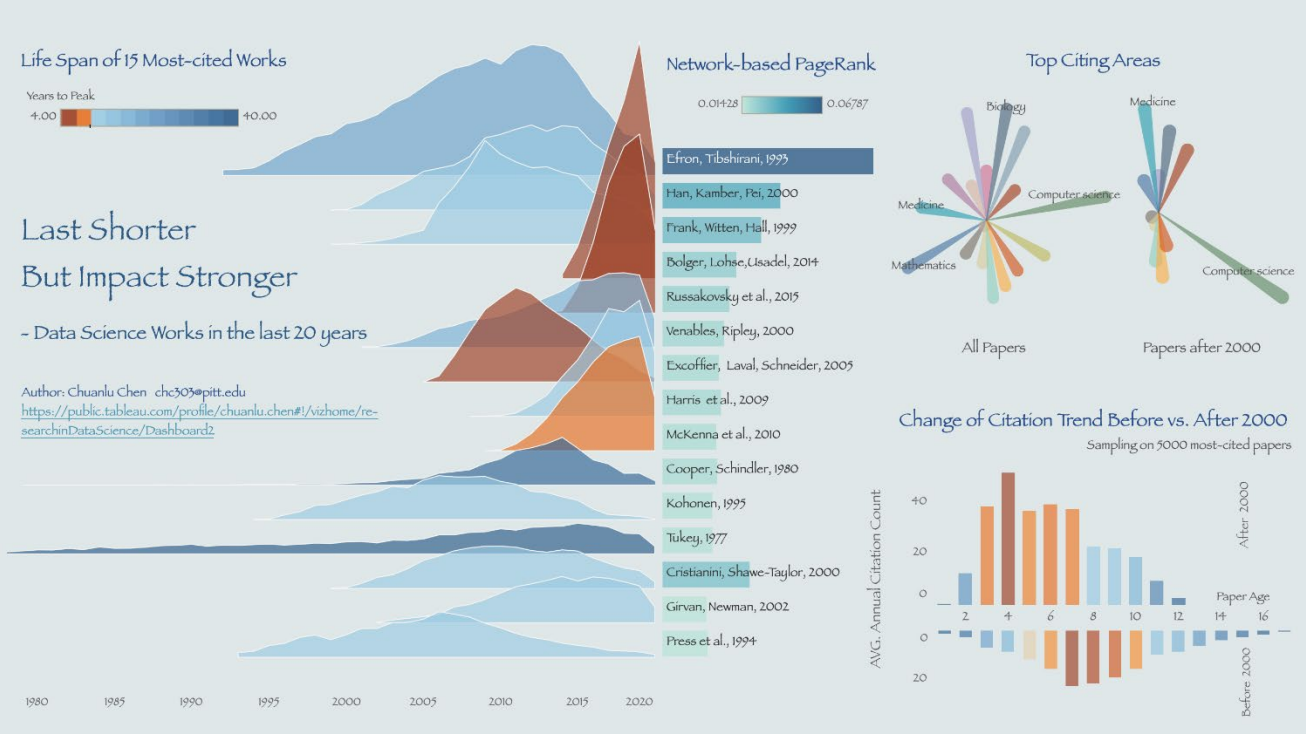
Chuanlu Chen SoftPeople ID: 4342604
Information Visualization - INFSCI 2415
Nov 28, 2020

Final Project Report

Finished Work



Legend

This project is to look at the research impact of papers in Data Science area. It has one main figure and three sub-figures.

The main figure (left: Life Span of 15 Most-Cited Works) shows the citation trend of the 15 most-cited papers in the area of Data Science. As a matter of fact, it indicates how long a research paper contribute to its academic field in terms of time. Therefore, I define it as 'life span' of a paper. In the figure above, the x-axis is time series, ranging from 1978 to 2020, and year is chosen as the granularity. And the y-axis is how many citations one paper gets in a single year. Specifically speaking, this is a normalized value. I preprocess the data beforehand to make them in a same scale. Those 15 papers are arranged in a descending order of total citation number from top to down. They are also colored in the order of how many years they need to get to the peak. The orange-red color palette shows the one paper needs less than 10 years or 5 years to reach the peak of its citation trend. If it needs more than 10 years to get reach its peak, it is shown in blue color. The longer, the darker. The ridgelines show the magnitude of research impact of a paper, as well as its rises and falls. Details about paper title, accumulative citation count and yearly citation number can be accessed via interactive tooltips.

Sub-figure No.1 (bar chart in the middle) shows the PageRank scores of each paper. PageRank is a linkage analysis algorithm used by Google Search to rank web pages, which introduces a node centrality metrics to

measure the importance of different nodes on basis of network. Complementary to citation-based metrics, this bar chart with PageRank scores provides us another perspective to evaluate the quality of research impact.

Sub-figure No.2 (pedal charts in the upper right corner) shows the distribution of study areas of those 15 papers. In essence, it indicates that where does the research impact go. I split this figure into two views. The one in the left shows the overall distribution of study areas of all 15 papers. In the right, I only keep the cited paper that published after 2000. Pedal length suggests the normalized average citation count in an individual area.

Sub-figure No.3 (dual bar chart in the lower right corner) makes a comparison between citation trend of paper published before and after the year of 2000. X-axis is the paper age. Data on the y-axis is annual citation count on average. Bars above x-axis belong to the paper published after the year of 2000. Accordingly, bars below x-axis are papers published before 2000. The average annual citation counts are computed based on the 100 papers sampling from the dataset. Bars with higher citations counts are highlighted by orange and red colors.

Significance

Scholarly articles are the coin of the realm in modern academia; they influence researchers' career paths, salaries, and reputations. The question is how to measure the impact of a research paper. One of the well-developed metrics is citation count, which is being used not only to measure the visibility, impact, and quality of articles but also to measure the performance of researchers. However, merely citation count is not able to explain everything in details. In this project, I would like to look at the research impact in details and in a couple of new perspectives, including the how long does it last, how important the impact is, and where does it go. More importantly, I would like to figure out is there any hidden pattern existing in these dimensions. To limit the scope of this project, I focus on the papers from the area of Data Science. By analyzing the paper life span, network-based page-rank score, and study areas of citing paper, the visualization suggests that papers in the 20 years tend to have a shorter life span and a stronger impact. In addition, the change of citation trends seems in line with the changing distribution of citing areas. These findings help us gain a better understanding of Data Science area.

Findings

1. Two distinct patterns of life span

The main figure helps us understand and quantify how article citations evolve as time goes by. Generally speaking, life cycle of all papers follows a similar path. In the first years after publication, articles generally receive a small but growing number of citations until, eventually, they reach a peak from which they then decline. However, this kind of life cycle trend seems to vary greatly in the emerging area of Data Science. Papers that published more than 20 years tend to be long-lived, gently-growing. They take relatively long time to reach to the peak, and contribute to the Data Science area gradually. For instance, impact of some papers (Turkey, 1977 and Cooper & Schindler, 1980) may keep in relatively moderate magnitude, but they last more than 40 years to exhaust their energy. On the contrary, papers that published in the last 20 years seem to have shorter life span, but leave stronger impact. For instance, the paper written by Lohse, Bolger, and Usadel in 2014 surged to unprecedented peaks (annual citation of 3450) in a period as short as 5 years. Then, it has a sharp decline. Other paper published after 2000 also follow the similar patterns. In a nutshell,

we can clearly see two patterns of life span from the main figure: one is a long-lived, gently growing pattern, and the other last shorter but leave a strong impact.

2. On average, papers after 2000 have a shorter life span, stronger impact and reach a peak rapidly

Our discoveries from the main figure have been further illustrated by the dual bar chart in sub-figure No. 3. From the 5000 most-cited papers, I collect 100 Data science papers that published before 2000, and 100 papers published after 2000. Then, I map the average annual citation count with respect to their paper age separately. The bar chart above x-axis shows the average peak of papers after 2000 is around 4 years with a mean value of 51.42. The average life span of papers after 2000 is about 12 years. On the contrary, the peak of papers before 2000 is around 7 years with a mean value of 19.09. For papers before 2000, the average life span is about 17 years. Observations from this figure has confirmed what we get from the main figure. Although the year of 2000 may not be a clear-cut boundary, we can see the overall trend changed around that period of time.

3. A longer life span does not necessarily mean a high-quality impact

The bar chart in sub-figure No.1 provides another way to look at the story. The top 1 paper are confirmed by its high PageRank scores. Additionally, three papers published around the year of 2000, including Cristianini, Shawe-Taylor in 2000, Han, Kamber and Pei in 2000, and Frank, Witten and Hall in 1999, achieve high PageRank scores, which suggest that they tend to be cited by more important papers and play important roles in the academic ecology of Data Science. However, I also find that some papers in the middle, though contributing to this area in a long term, do not have high PageRank scores, which means that they tend to be cited by less important works. Their research impact may not be compatible with others among this top 15 paper list.

4. Papers after 2000 contribute more to its own area

The pedal charts in sub-figure No.2 demonstrate where does the research impact go. Each pedal suggests a citing area. Pedal length indicates the normalized average citation count in a specific area. Generally speaking, papers in the area of Data Science contribute to a variety of fields. To name a few, the top citing areas are Mathematics, Biology, Business, Medicine, Psychology, beside Computer Science which Data Science belong to. However, for papers after 2000, the distribution of citing areas is extremely skewed. A vast majority of the citations are coming from the its own area. Only a small amount of the research impact goes to other fields. It seems the circulation of research impact mostly happened within its area.

Data and Method

The major data resource for this project is Microsoft Academic Graph. This visualization project involves several datasets from this API, including Papers, Authors, Paper Author Affiliations, Paper Fields of Study, Field of Study, Field of Study Children, Field of Study Extended Attributes and Paper References. I use paper fields of study as an attribute to filter the data, since I only focus on the Data science area in this project. The total number of papers in Data Science area is 425742, and the overall citation count is 388,3836.

As for the main figure, I retrieve the top 15 most-cited papers in the Data Science area. Then, I find the papers that cited those 15 papers from the MAG API. Thus, I construct a citing paper dataset in the size of 161,0839 samples. It is surprising that the citation count of the top 15 papers in Data Science covers almost 40% of the overall citation count. The next step is really straightforward. For those top 15 paper, I count how many citations a paper has received in each year, and visualize them in a ridgeline plot.

As for the sub-figure No.1, I construct a network on basis of the citing paper dataset I created before via Python networkx library. Then, I use the build-it function to compute PageRank scores for each of the 15 papers. Consequently, I demonstrate the results via a bar chart.
As for the sub-figure No.2, the distribution of citing areas for each paper has been mapped on the pedal chart. While an individual paper being selected, the pedal chart in the left shows the distribution of citing areas for the single paper. Otherwise, it shows the overall distribution for all 15 papers. To make a comparison, the pedal chart in the right shows the citing area distribution for papers after 2000.
For the sub-figure No.3, first of all, I filter the 5000 most-cited papers from the Data Science paper dataset. The reason to do this is that most of the papers do not have any citation. This kind of paper does not help us to explore citation trends. Then, I do some sampling on the filtered data. I retrieve 100 papers published before 2000 and after 2000 separately. Based on these two samples, I compute how many citations a paper has received in each year on average since it was published. I plot the results in two bar charts, so we can compare between them easily.

Discussion

There may be several possible explanations for this phenomenon.
Firstly, Data science has experiencing a rapid growth and fast iterations in the last 20 years. In the recent years, academia has been making efforts to adjust to the growing demand for data science and data scientists. These include academic journals dedicated to data science and big data, as well as scientific meetings and conferences. The large demand for Data Science speeds up the process of idea generation, knowledge sharing and knowledge upgrade. That may explain why papers in Data Science tend to have a shorter life span but reach its peak quickly.
Secondly, although pattern recognition and machine learning are not new things, Data Science is very young. Data Science has grown to be an individual academic area in the last 20 years. In the very 20 years, Data Science constructs its own core values, internal cohesion, scholar community and research paradigm as an integrated discipline. In addition, the field of study attribute of MAG is generated by NLP techniques on paper context. Specifically speaking, earlier papers labeled as data science may not be the 'Real' Data Science papers but covering some basic ideas or knowledge relevant to Data Science. Therefore, they can be easily cited by other researches in areas.
In addition, with the development of Data Science area, researches go more detailed, professional and in-depth. That prompts research mostly circulate within its own area. Also, outstanding researches leave more profound impact than before.

CVPR analysis

Haonan Duan
Contact: HAD65@pitt.edu
The University of Pittsburgh

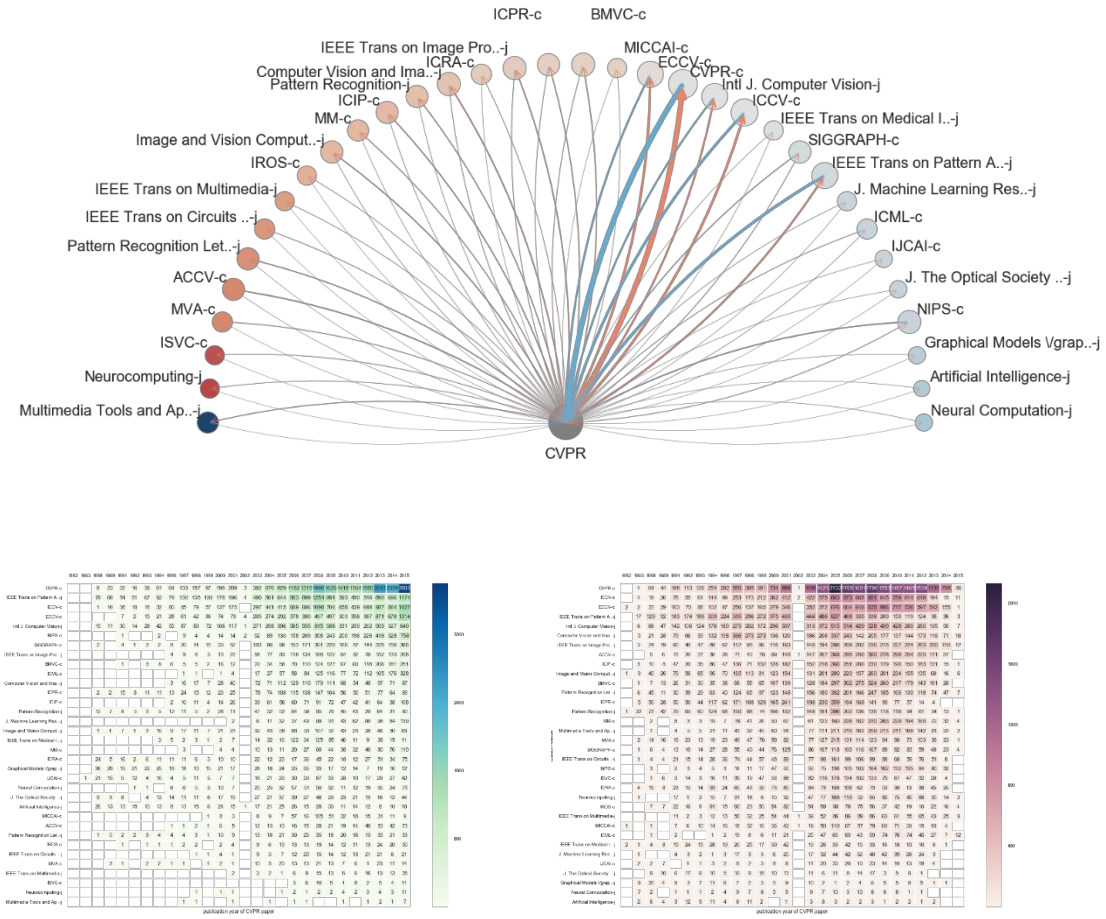


Figure 1. CVPR analysis. The datasets used will be discussed for more details in Data section. The main source of data is MAG, which include 1283 and 23404 types of conferences and journals correspondingly. **a,** Summary of incoming vs outgoing citations to the top-50 venues. I set top K to be 25 for either references and citations, the final plotted conferences and journals number are 34 since some papers appear in both references and citations. Node colors: ratio of citations (outgoing, red) vs references (incoming, blue). Node sizes: amount of total citations and references in either direction. Edge thickness: blue edges are scales by the number of references going to a given venue, red edges are scaled by the number of citations coming from a given venue. Nodes are sorted left-to-right by the ratio of incoming vs outgoing citations to CVPR. **b,** Heatmap of references over time. The contribution to CVPR of top-50 (actual selected are 34) papers are quantified. Along with the y-axis, the contribution of each conference or journal to CVPR increases with the axis increases. **c,** Heatmap of citations over time. The interest to CVPR of top-50 (actual selected are 34) papers. The conference or journal is more interested of CVPR researches and topics with the y-axis increases. Figure1 mainly shows the role of CVPR in computer science area and its historical tendency. Majority of conclusions in the report are obtained from this.

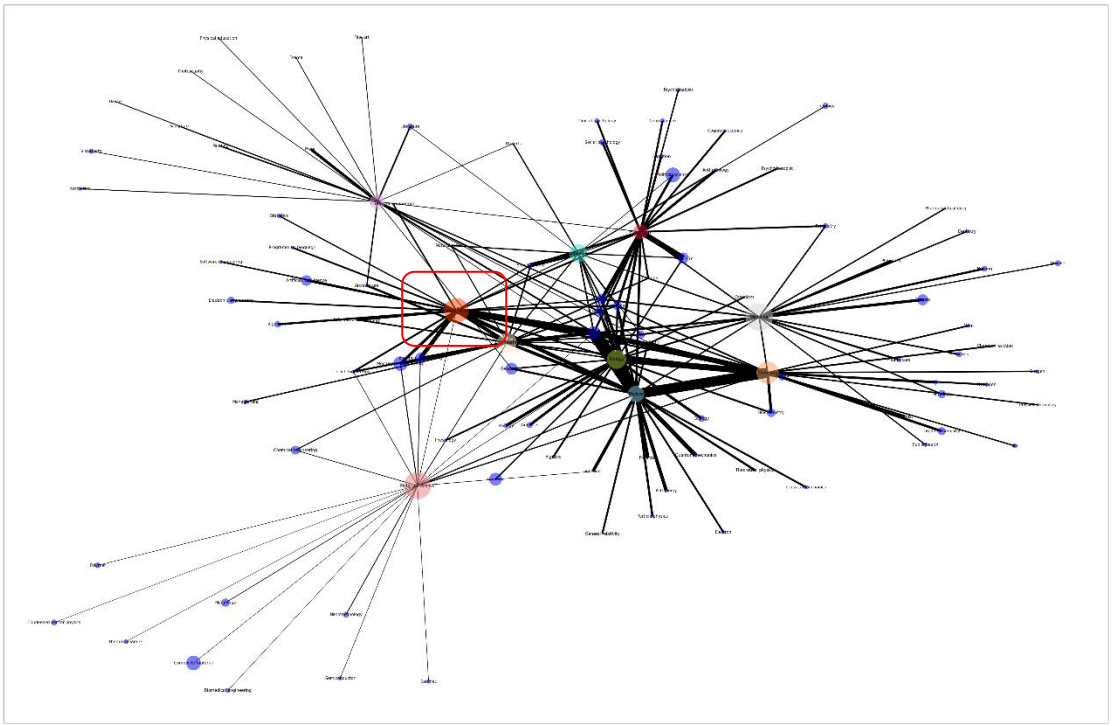


Figure 2. The related connection between computer science and other fields. The figure is plotted from table *FieldOfStudy* and *RelatedFiledOfStudy*. FieldOfStudy is sorted by *PaperCount* attribute and top-10 fields are selected. Node size: The related amounts of papers in this field. The computer science node is highlighted by a red rectangle. As we can see, computer science field plays a very important role in study and can be applied to numerous fields. In addition, by comparing the node size, computer science has a large number of papers among many fields.

Figure 3. The fields cloud. By counting the frequency of fields name, I obtain a fields cloud with name of each field. The larger the font is, the more frequencies it appears. Computer science is a relatively large font in the cloud, which indicates its influence and researchers' interests on it.

Significance

This project responds to the analysis of one of the most popular computer science conference – CVPR. Motivated by the growth of computer vision and its applications make people's life increasingly convenient, I explore the how quickly CVPR develops along the time. By analyzing nearly 30 years references and citations in each CVPR paper, I find CVPR is seen an explosive growth after the first deep convolutional neural network proposed. I also demonstrate that with the progress of computer vision techniques, they attract more and more interests in other communities. Also, nowadays, majority of computer vision tasks are accomplished by deep learning, some conferences and journals in traditional fields are decreasingly cited by CVPR.

Findings

1. *CVPR has become one of the best conferences.* In the summery figure, nodes are sorted left-to-right by the ratio of incoming vs outgoing citations to CVPR. As we can see, CVPR locates center-right and with nearly white color node, which indicates CVPR has achieved a balance of incoming and outgoing. This balance is a very important sign for a conference or journal since most of conferences or journals has more incomings than outgoings. The conference or journal can be said acquire attentions from researchers once it obtains this balance.
2. *CVPR topics become more diverse.* In the heatmap of references over time, I find CVPR cites more and more papers mainly focus on other fields. Such as ICRA and IROS, these two are top conferences in robotic community, which are cited by CVPR with an increasing number. The tendency exhibits that CVPR begins to transform from concentrating on computer vision theories and models to interdisciplinary applications.
3. *Interactivities among computer vision community grow rapidly.* In the heatmap of citations over time, it shows the color of top-right corner get deeper and deeper. By checking the names, almost all of them are the conferences and journals mainly focus on computer vision, such as ICCV, ECCV, Trans on PAMI, IJCV. The color change indicates that the interactivities among computer vision community, especially those best conferences and journals grow rapidly.

Data

In the MAG, used tables are listed below:

Papers: the brief of each papers in dataset

ConferencesSeries: the paper will be linked to this table for more information if it's from conference

Journals: the paper will be linked to this table for more information if it's from journal

PaperReferences: the references of the paper

PaperCitationContext: the citations of the paper

Method

For each paper in papers table, check if this paper is a conference or journal by *JournalId* and *ConferenceSeriesId*, and also the *Year* of this paper published. From the obtained *JournalId* or *ConferenceSeriesId*, refer to *Journals* or *ConferencesSeries* table to get the *NormalizedName* of the conference or journal. Combining the *PaperId*, *PaperReferences* and *PaperCitationContext*,

the references and citations information of each paper will be acquired. A dictionary is used to store the references and citations amount for each conference and journal year by year. *Networkx* module is used to accomplish the visualization.

Discussion

With the development of computing hardware, such as GPU, numerous theories proposed in last century are able to be implemented by modern computers. Deep learning, one of the most powerful methods that enhances quality of people's left based on the GPU. Computer vision is one of the most exciting deep learning applications, makes many fictions in novels or science fiction movies come true. The potentials of computer vision also attract increasingly number of researchers. On basis of this background, I analyze one of the best conferences in computer vision community – CVPR.

After several-year growth, CVPR has become one of the best conferences. Even its influence cannot come up to NIPS, ICML, etc., which are like the ceiling of computer science papers, CVPR still gain the balance of references and citations. This indicates that the research works that are able to be published on CVPR are sloid and authoritative, which can inspire other researchers many ideas. And with the computer science area rapidly develop, interdisciplinary applications become more attractive than computer vision theories itself. Computer vision techniques can be deployed in many tasks, such as intelligent robots and autonomous cars. Results from this, diverse conferences or journals from other communities come into the references of CVPR. In addition, interactivity of acknowledged top-3 computer vison conferences, CVPR, ICCV and ECCV is tighter.

There are still a lot of rooms for growth of CVPR, and by analyzing the tendency of CVPR over nearly 30 years, I believe it will finally become a best conference like NIPS and ICML.

Field Flowers: Visualizing Field Citation Patterns

Zak Risha zak.risha@pitt.edu

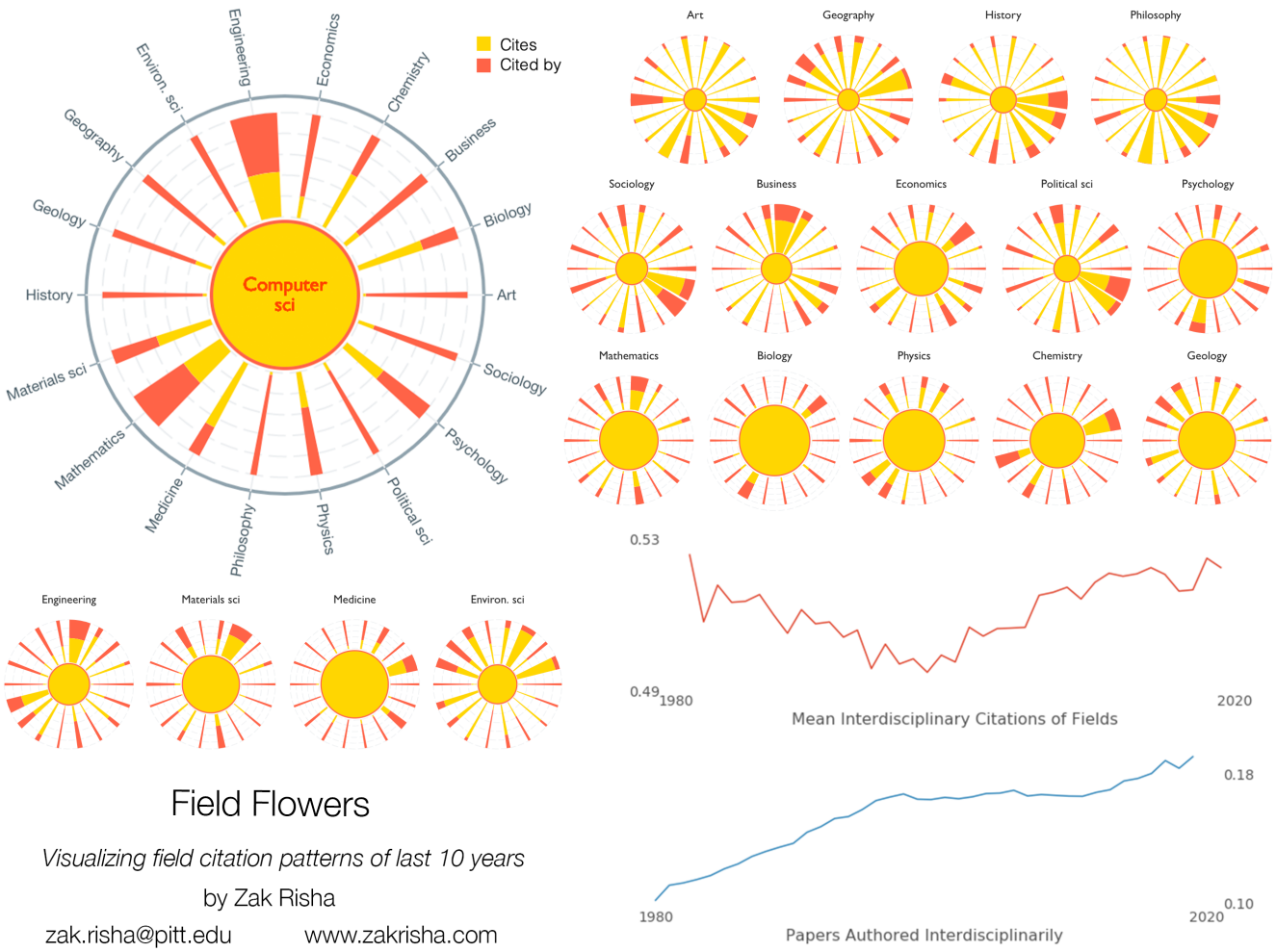


Figure 1. Field Flowers: Visualizing Field Citation Patterns. This figure introduces a new visualization to display the citation patterns of a specific field, more specifically how they interact with other academic disciplines. The Microsoft Academic Graph (MAG) powers these figures, consisting of 907,021,934 citations within the last decade. In addition, 181,277,688 papers and 1,597,190,739 citations are visualized in the summary graphs which cover since 1980. MAG implements a 6 level topic hierarchy for their dataset with 19 top level nodes representing the prominent academic disciplines. All papers are situated in this hierarchy, allowing relations such as citations to be analyzed and visualized based on the labeled field of study. Each field is visualized as a flower with the pistil representing how frequently the field cites papers from within its own field. This is an important aspect to highlight in the context of interdisciplinary research, and the size of the center gives an immediate and intuitive shape to a field revealing the level of insularity from other disciplines. Each field branches out to all other disciplines creating petals, with the width revealing the degree to which a field cites that other discipline. This redistributes area to proportionally alter the shape of the flower. In addition, the colors of each petal show the balance of citations. Specifically, the gold represents the proportion the field cites the other and the red represents a reciprocal relationship. If petals are overwhelmingly red like in computer science, it means a larger percentage of papers in other fields cite computer science than are cited in return. Therefore, more interdisciplinary fields typically have smaller centers with thicker, gold petals and more intradisciplinary fields have larger centers with skinnier, red petals. The first summary graphs simply shows how all fields' level of interdisciplinary citations (thickness of petals) have fluctuated over time. The second tracks the percentage of papers with interdisciplinary authors by calculating each author's field of study based on which field they have published the most papers.

Significance

Interdisciplinary research has become a buzz word in academia to describe research, schools, students, curriculum, and faculty. In the past 25 years, this aspect of knowledge has become valued by administrations for the ability to handle problems one discipline alone cannot tackle, resulting in novel ideas. Initiatives in the past 25 years have sought to grow this type of research, such as the Department of Defense MURI Grant that began in the late 90s and can award as much as 200 million dollars in a given a year. Similarly, collaborative and interdisciplinary grant opportunities have also sprung up from the NSF. Given these changes, visualizing interdisciplinary research's current state and changes over time gives us crucial insight in how the structure of academic knowledge and production is changing. Patterns in research publications and citations can be used as a rough proxy to calculate interdisciplinarity, and the Microsoft Academic Graph has already made some progress. Microsoft itself has generated some visualizations based off this information, most notably a citation matrix[1]. However, such a graphic presents a substantial amount of information that is hard to immediately interpret. A metaphor and visualization like field flowers can more immediately and intuitively present the characteristics of a specific field.

Findings

1) *Disparity among fields and groups of fields*. What's immediately apparent is that each field is unique in its composition and reference of knowledge. Overall knowledge flows unevenly, leading us to question how differences in objectives and research methods of fields contribute to transfer of knowledge. Flowers are clustered together with similar fields to help visualize differences among related disciplines. The fields in the humanities tend to be the most interdisciplinary, as they more frequently reference one another and other disciplines. Fields grow somewhat more intradisciplinary in the social sciences, which often build off the humanities but introduce their own methods in more applied research. Unlike the humanities, the hard sciences constitute the most insular of fields, sharing similarly large pistils. As fields become more applied, like in the case of Engineering, they may grow more interdisciplinary likely in efforts to tackle diverse problems. However exceptions, like medicine, grow increasingly insular.

2) *Interdisciplinary research has decreased, but now is increasing.* By looking at citation trends of fields over the last 40 years, we can see the 90s were actually the start of growing intradisciplinary as self referential citations grew to new heights into the early 2000s. However, possibly due to some of the initiatives previously described, the trend is reversing and interdisciplinary citations have rose and all fields are averaging over 52% interdisciplinary citations.

3) *Interdisciplinary authorship has increased.* Our analysis of the number of papers with authors from differing fields of study reveal a very strong trend in academia. Papers are increasingly being produced by interdisciplinary teams with over an 8% increase over the last 40 years.

Data

Microsoft Academic Graph (MAG). MAG is a massive academic dataset that contains academic papers, citation relationships between them, and additional information such as authors, fields of study, and affiliations. Details about the dataset and access are available at *https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/* . The dataset can be accessed over Microsoft Azure freely for new accounts.

---

[1]https://academic.microsoft.com/topics/

Method

As mentioned, the data visualizations are built on MAG's ontology of fields. From there, citations can be calculated to papers labeled with specific top level fields. This data was then processed by year, calculating the citations to a field from all papers in a particular field, in a particular year. The last decade was used to generate these visualizations after converting the data into an unstructured format (JSON). This visualization was constructed with the React library Victory which is a web based system that relies on SVG to render chart components. Victory supports polar (rose) charts and allows for configuration of many subcomponents. The metaphor used was a flower, which was used previously by Shin et al. to visualize more fine grain influence among authors and institutions[1]. For field flowers, modification to components manipulate the shape of these flowers to reflect citation patterns and interdisciplinary research.

Discussion

The field flowers reveal distinct disparity among the fields located within the humanities and social sciences versus hard sciences. There are a few notable explanations to citations patterns the visualizations revealed. One could be related to the objective of these fields differ dramatically. Disciplines that are more theoretical like the humanities and sometime the social sciences may prioritize asking questions or identifying problems, rather than proposing concrete solutions. Even if problems are addressed, they may be theoretical. These disciplines might cite more applied fields to critique or further inquire about implications of solutions. However, hard sciences and applied sciences may be more invested in tangible solutions, preferring to cite previous related attempts that detail a specific method, approach, or result.

Similarly, another contributing factor to disparity could be differing research methods. Within some disciplines, empiricism takes utmost precedence. Hard sciences and applied sciences often have idiosyncratic research conventions and protocols that limit the ability to incorporate other disciplines. A great example is medicine, which is the most insular field but also has the most at stake. When human lives are being considered, non-empirical or speculative research would likely face harsh criticism and fail to meet the standards embodied by the field.

When addressing whether interdisciplinary research is increasing, there are conflicting results between authorship and citation. While interdisciplinary authorship shows steady increase, citation reveal a differing pattern of decline, then increase. It could be that while a paper was authored by scholars from differing fields of study, the resulting research failed to impact more than a single field. This could also be very concerning, meaning that interdisciplinary teams try to coerce their research to one of the disciplines so that it will align with the existing landscape of academic research. This could indicate a need for more interdisciplinary venues that can navigate the varying standards of multiple fields.

There are some limitations to this research as conclusions are driven by the composition of the MAG dataset and field of study hierarchy. Disparities found might be related to what papers are included in the dataset. Perhaps there is bias towards fields in hard and applied sciences, and thus the dataset includes more interdisciplinary papers in the humanities that could relate to these fields. Additionally, as the process of labeling paper's field of study is automated, these findings are dependent on the accuracy of the algorithm.

References

1. Minjeong Shin, Alexander Soen, Benjamin T. Readshaw, Stephen M. Blackburn, Mitchell Whitelaw, and Lexing Xie. 2019. Influence Flowers of Academic Entities. In 2019 IEEE Conference on Visual Analytics Science and Technology (VAST), 1–10. https://doi.org/10.1109/VAST47406.2019.8986934

The Academic Development of the Payment Field

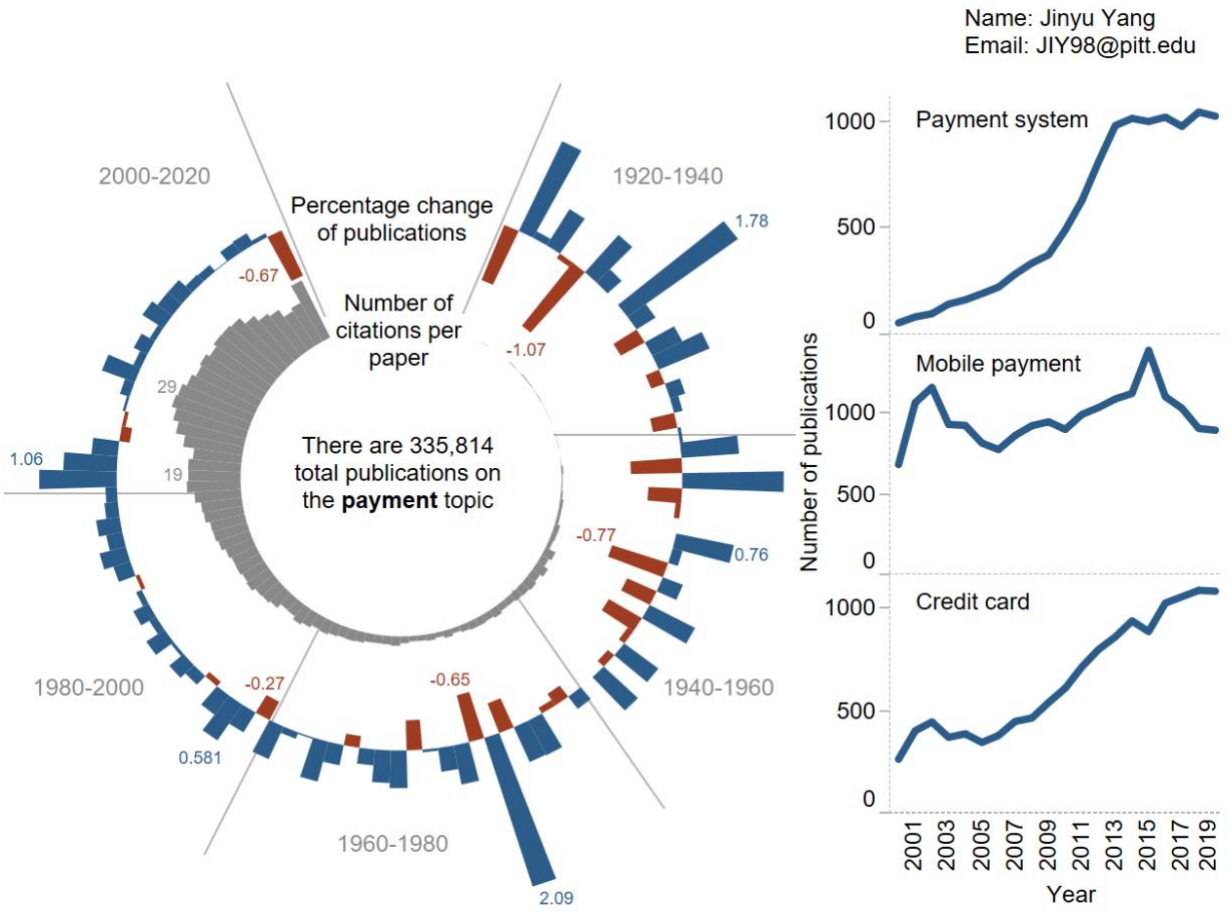


Figure 1. The Academic Development of the Payment Field. This set of graphs are built on one main dataset – Microsoft Academic Graph, which includes academic records about various fields of study. **a,** Visualizing the change of publications and citation per paper. The graph starts from the upper left and moves clockwise to show annual data from 1920 to 2020. It consists of three parts. Firstly, the outer circle shows the percentage change of publications for each year. Blue means positive rates of change, and red means negative. The circle is separated into five sections, with 10 years in each section. The corresponding timeframe is labeled next to the circle. Within each section, the maximum and minimum values are labeled for reference. The height of the bar is proportional to the rate of change. Secondly, the inner circle shows the number of citations per paper published each year. The value will only be positive. It follows the same segmentation of years as the outer circle. Because the purpose of the inner graph is to show the trend over time, instead of value, I labeled two points – 19 and 29 to show reference. The height of the bar is proportional to the value of citations per paper. Thirdly, the sentence in the center gives information about the total amount of publications in the payment field of study so far. **b,** The trends of child field of study are different. This graph consists of three parts, corresponding to three main child fields of study under payment: payment system, mobile payment, and credit card. Three line graphs are arranged by the number of publications in 2019 in descending order from top to bottom. Each line graph tells the number of publications for each year from 2000 to 2019. To estimate the number of publications in the next year based on the trend, payment system is likely to have the same amount of publications, mobile payment is likely to have fewer publications, and credit card is likely to have more publications.

Significance

This project helps readers to understand the academic development of the payment field of study. In the preliminary research, I found that payment has a much higher z-score in terms of the number of publications for each year within the past 5 years. To figure out why, I visualize the history of development – when it started, how the number of publications progressed through time, how deep authors interact with each other (by calculating the number of citations per paper). In addition, I visualize the recent activities of three main subtopics, which can give a hint about what this field includes and what direction it will move in the future. For people who are curious about this field, this visualization will be a good start point to gain a comprehensive understanding, then zoom in on a particular segment of time according to their interests.

Findings

1. *The payment field of study has a stable percentage change of publications throughout the past 10 years*. The rate of change has a small peak around 2000 (with a value of 1.06). After that, the rate has small variations, which indicates that this field has passed the exponential growth stage, moving towards a more sophisticated stage. In addition, this field has a long history – started in 1920.

2. *The number of citations per paper seems to have a cycle over time*. From 2000 to 2010, the number of citations per paper increased over time, from 19 to around 29. Then the value gradually decreased. From the data for the most recent year, it looks like the value is increasing again.

3. *The credit card sub-field will likely have more publications in the next year*. By examining the activities of three main sub-fields, we can see a clear trend that payment system has saturated, mobile payment is likely to have less attention, and credit card is likely to keep having more related papers published.

Data

Microsoft Academic Graph (MAG). MAG is an online database that contains information about academic fields of study, related publications, authors, institutions, journals, and conference. This project mainly uses data related to papers.

Method

From a large amount of available data, I made a delicate choice of measures to present and format of visualization.

To show the development history of this field, I choose two measures: percentage change of publications and citations per paper. The calculations are below:

$$Percentage\ Change = \frac{P_t - P_{t-1}}{P_{t-1}}\ (X = number\ of\ publications, t = year)$$

$$Citations\ per\ paper = \frac{P_t}{C_t}\ \ (P = number\ of\ publications, C = number\ of\ ciataions, t = year)$$

For the number of publications, I choose to show the percentage change instead of the original value, because the original value is likely to grow over time, and it will be hard for readers to tell the difference, given the height of the bar is small.

For citations per paper, I didn't use the cumulative value because I believe each year should be considered individually. I didn't use the original value because it's hard to tell whether the change in citations is due to the change in the number of papers. So I use the ratio to find out how authors are interacting with each other through time.

For the timeframe, I choose to show complete data. From 1920, the first finance paper got published until now. It will give readers complete information and avoid misleading. Because it includes so many years, presenting it in a horizontal bar chart is no longer ideal. So I curved the x-axis to a circle.

For the panel on the side, I choose to use the original value because I only want to show a few years of data in a very straightforward. I also included it to give some hints to the reader about what the payment field is about.

Discussion

At the beginning of this project, I have several hypotheses, with most of them proved wrong during my exploration.

Starting with preliminary research, I calculated the number of papers for each subtopic of finance for each year, then pick the top 5. Secondly, I calculated the z-score for each one using this equation:

$$z_{a,b} = \frac{x_{a,b} - \mu_a}{\sigma_a} \qquad a = Year; b = Field\ of\ study$$

I thought the z-scores will be similar and very small, given finance is a very sophisticated area of study. But the results show that the z-score for payment is around 1.4, while other areas have absolute values smaller than 0.5. It is worth noting that the standard deviation of numbers of publications for each year is very large, because finance has many sub-topics and some of them have single digit number of paper published per year.

So I decided to zoom in on the payment field of study and visualize its progress over time.

In terms of the circle, if we split it in the middle, the left half tells more information than the right half, because the field started to grow around 2000. I found that 80% of the papers are from 2002 to 2019 (17 years), and the left 20% are from 1922 to 2002 (80 years).

In addition, my hypothesis was that since the number of papers grows over time, citations per paper will increase. But this idea was proven wrong according to the graph. In fact, we observe a small hill on the left half of the circle. The number of citations per paper started to grow gradually from 2000, reached a peak around 2010, then decrease.

Zoom in further to the sub-fields of finance, I choose to show the top 3 sub-topics (ranked by the number of papers published in 2019). We can see that the payment system and credit card both went through large growth from 2000 to 2019. But mobile payment seems to have a mixed trend. From the trends, it seems that the payment system has entered the mature stage, mobile payment is still developing, and credit card is continuing to grow.

Rank of Top National Universities Across Research and Other Categories

Jaime Allen Fawcett (jaf157)

Significance and Background

The college application process can be daunting for students and families alike. There is a wealth of information available online to highlight various attributes for different universities and colleges, such as academics, student/teacher ratio, cost and other financial considerations, student life, and others. While several online resources such as US News World and Report, Niche, and The World University Rankings offer consolidated information about various parameters or domains, all of these outlets rely on the *league table or ranked list* – a scrollable list that allows the user to view 'snapshot' information about universities in order from best to worst.

Ranking and league tables have become important tools for university and college comparison in a few ways. First, as we discuss, they are useful in allowing users to easily view a university's or college's standing across different parameters, whether it is best budget, best student life, or best academics[1]. They are also critical to helping universities and colleges establish prestige and reputation[2], not just overall in a nation, but also amongst similar schools that would participate in the same market or advertise to the same student base. After all, it is not always a student's desire to attend an Ivy League college, so a list titled "Best State Universities" may hold greater appeal but feature very different universities.

However, from an information processesing and usability standpoint, these ranking and league tables have limitations. These limitations are associated with "display fragmentation" which occurs when information needed to complete a task/decision is spread across different screens or webpages rather than located on the same screen. This presents itself in a phenomenon known as the *keyhole effect* where only limited information is viewable on a single screen, similar to looking at a large room through a keyhole. For example, these websites feature several different lists that have to be viewed individually – overall ranking is a separate list from cost, which is a separate list from rankings on innovation or student life. Additionally, these lists typically only allow the user to view a small number of universities on the screen at one time (2-4 at a time for a list that can span hundreds). Figure 2 highlights these limitations; on three university ranking websites, at most only two to four universities on the list were viewable at any one point. While the "cards" used in the list feature different information, the ranking is limited to one domain: overall ranking.

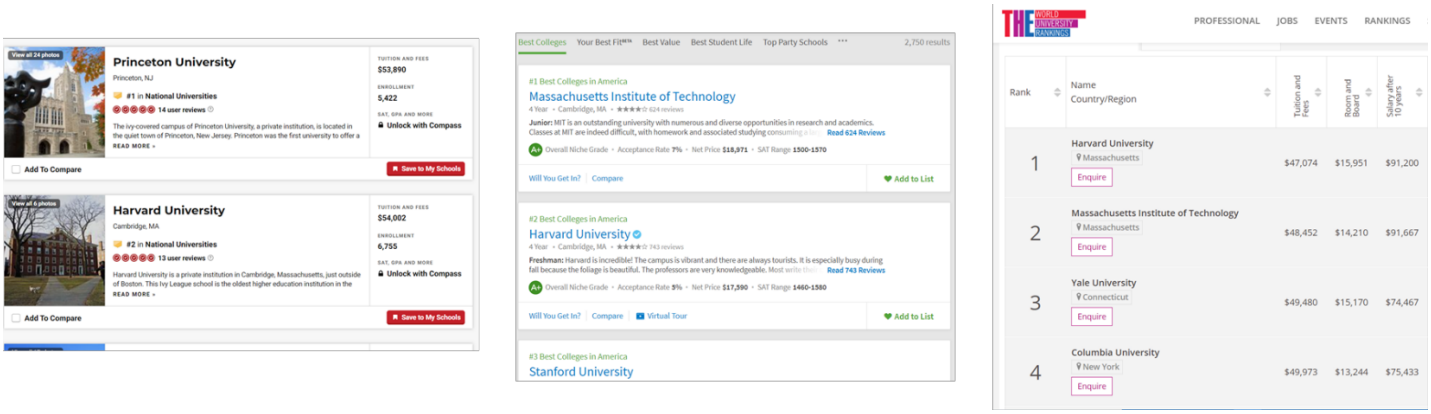


Figure 1. Ranking and league tables from three university ranking websites. On each website, only one list is viewable on the page/screenand only 2 – 4 universities or colleges are viewable at a time.

Display and information fragmentation has consequences for usability and cognitive load. First, because the information is spread out across a scrollable list or several scrollable lists, users are required to hold information in memory in order to make comparisons. A student comparing two universities across six lists, must navigate to each list, identify each university, identify their ranking, hold those two numbers in mind, and then do the same on five other lists; thus, the studen it required to hold 12 different numbers in mind and remember which sets of numbers apply to which university. This places strain on

[1] Drewes, T., & Michael, C. (2006). How do students choose a university?: an analysis of applications to universities in Ontario, Canada. Research in Higher Education, 47(7), 781-800.

[2] Hazelkorn, E. (2007). The impact of league tables and ranking systems on higher education decision making. Higher education management and policy, 19(2), 1-24.

working memory, which we know is limited, and thus quickly makes the task of researching universities taxing (increases cognitive load). Another strategy users might take is externalizing the information by aggregating and recording it in a separate spreadsheet. Again, this places the onus of work on the user.

Therefore, in order to improve the university/college research process and minimize cognitive load and burden on the user (the student or family members), this project aimed to create a new visualization utilizing a bump chart that would allow easy visualization and comparison of several universities (up to 25) across several domains (up to 7) at once.

Methods

The following describes the data and methods used to analyze and create the visualization.

Datasets

Data used for analysis and visualization were pulled from various sources, including Microsoft Academic Graph, US News and World Report, Niche.com, and PayScale.com. US News World and Report, Niche.com, and PayScale.com feature already established ranking lists. Wherever possible, data for the academic year of 2020 – 2021 were used. Data from these sources were used or extracted to create 7 different variables for ranking the sample universities.

Sample

A sample of 25 United States universities and colleges were selected. These universities were the 25 top universities as determined by US News and World Report 2021 college rankings.[3]

Variables/Domains

A total of 7 different variables were created:

1. Overall – A university's or college's overall ranking as indicated in US News and World Report.
2. Academics – A university's or college's ranking in terms of academic rigor as indicated by Niche.com.
3. Innovation – A university's or college's ranking in terms of how innovative they are as indicated by US News and World Report.
4. Research – A university's or college's ranking in terms of research prominence as indicated by h-index calcuated from Microsoft Academic Graph data.
5. Cost – A university's or college's ranking in terms of annual tuition cost as indicated by US News and World Report.
6. Salary – A university's or college's ranking in terms of average salary by alumni as indicated by Payscale.com.
7. Return on Investment (ROI) – A university's or college's ranking in terms of return on investment based on cost and salary and potential salary over time as determined by Payscale.com.

These domains can be split into two broad categories: Financial (Cost, Salary, ROI) and Academic (Academics, Innovation, Research), with Overall representing a more wholistic construct that would, ideally, factor in all of the other variables or domains.

Analysis

Overall, academics, innovation, research, cost, salary, and ROI variables were all determined by either extracting the rank value as is from the already established ranking list, or extracting a raw value and calculating the rank after all raw values were pulled for the sample (e.g., cost is $59,000, rank is 6). If a school was not featured on a list or their rank could not be determined, there Overall rank was used as a supplement. For example, if 15 of the 25 schools were ranked on one list, to determine the 16th spot, the remaining 10 schools would be examined and the school with the highest overall rank would be the 16th spot.

Research domain rankings were determined using the Microsoft Academic Graph datasets which contain information on universities, authors, and associated publications. Specific tables used in the data analysis include "Affiliation", "Author", "Papers", and "PaperAuthorAffiliations". The data schema for these tables and datasets can be found: https://docs.microsoft.com/en-us/academic-services/graph/reference-data-schema. For each of the 25 universities, an

[3] https://www.usnews.com/best-colleges/rankings/national-universities

average h-index value was calculated. H-index is a quantitative measure based on the number of publications and citations an author has and is meant to provide "an estimate of the importance, significance, and broad impact of a scientist's cumulative research contributions."[4] University average H-index was calcuated by extracting the appropriate data for each university, then individual H-index was calculated for each author at each of the 25 universities. Next, the h-index was averaged across all authors for each university. Therefore, each university had an average h-index value representing their research impact with a higher h-index representing a higher research impact. This metric was used to determine university/college Research rankings.

Once all rankings were extracted and aggregated, some simple statistics were computed including range (on raw numbers where available), and correlations. The visualization was constructed manually.

Results

The average H-index across universities was 2.89, with a maximum of 3.67 (California Institute of Technology) and minimum of 2.20 (Georgetown University). Average salary across the sample was $87,427, with a maximum of $113,053 (California Institute of Technology) and minimum of $53,868 (Emory University). The average cost (per year) for the selected universities was $56,091, with a maximum of $64,380 (Columbia University) and minimum of $42,980 (University of California – Los Angeles). 2020 reports of average US tuition and fees were $21,184 per year for out of state public institutions, and $35,087 for private institutions. So, we can see that all of the universities in our sample are greatly above these average costs with even the lowest cost institution, University of California – Los Angeles, a public institution, costing more than the average cost for private institutions.

Correlations between domains are presented in Figure 2. Correlations between Overall rank and other domains were as follows: 0.51 with ROI, indicating a moderate positive correlation, 0.58 with Salary, also indicating a moderate positive correlation, and -0.18 with Cost indicating a slight negative correlation (better schools were more expensive). On the academic side, Overall rank was highly correlated with Academics (0.88), as expected. Overall rank was only slightly correlated with Research and Innovation, indicating that it is possible that these factors were not considered in Overall ranking. Correlations between the Academic domains were only slightly positively correlated.

	ROI	*Salary*	*Cost*	*Overall*	*Academics*	*Research*	*Innovation*
ROI	1.00						
Salary	0.79	1.00					
Cost	0.11	0.07	1.00				
Overall	0.51	0.58	-0.18	1.00			
Academics	0.47	0.59	-0.26	0.88	1.00		
Research	-0.07	0.06	-0.14	0.25	0.33	1.00	
Innovation	0.58	0.59	0.28	0.35	0.37	0.21	1.00

Figure 2. Correlation matrix for domains of interest.

While the correlations provide a quick overview of the relationship between the domains, which could provide insight into a university's standings among the different domains, the figure below – a comprehensive bump chart – allows easy discernment of a university's ranking among the different domains. We can see some of the correlations reflected in the slopes of the lines connecting the different university icons between domains. For example, the slopes of the lines between the Overall and Academics domains are relatively flat, with universities typically shifting only a few spots from one category to the other. This is reflective of the high correlation between these two categories. Conversely, Overall and Cost featured a slight negative correlation. This is reflective of the steep slopes of the lines between the universities in each of these categories. Several high ranking universities in the Overall category, drop to the bottom of the list for Cost, and several low ranking universities in the Overall category, jump to the top of the list for lowest Cost.

[4] Hirsch JE. An index to quantify an individual's scientific research output. Proc Natl Acad Sci U S A. 2005 November 15; 102(46): 16569–16572.

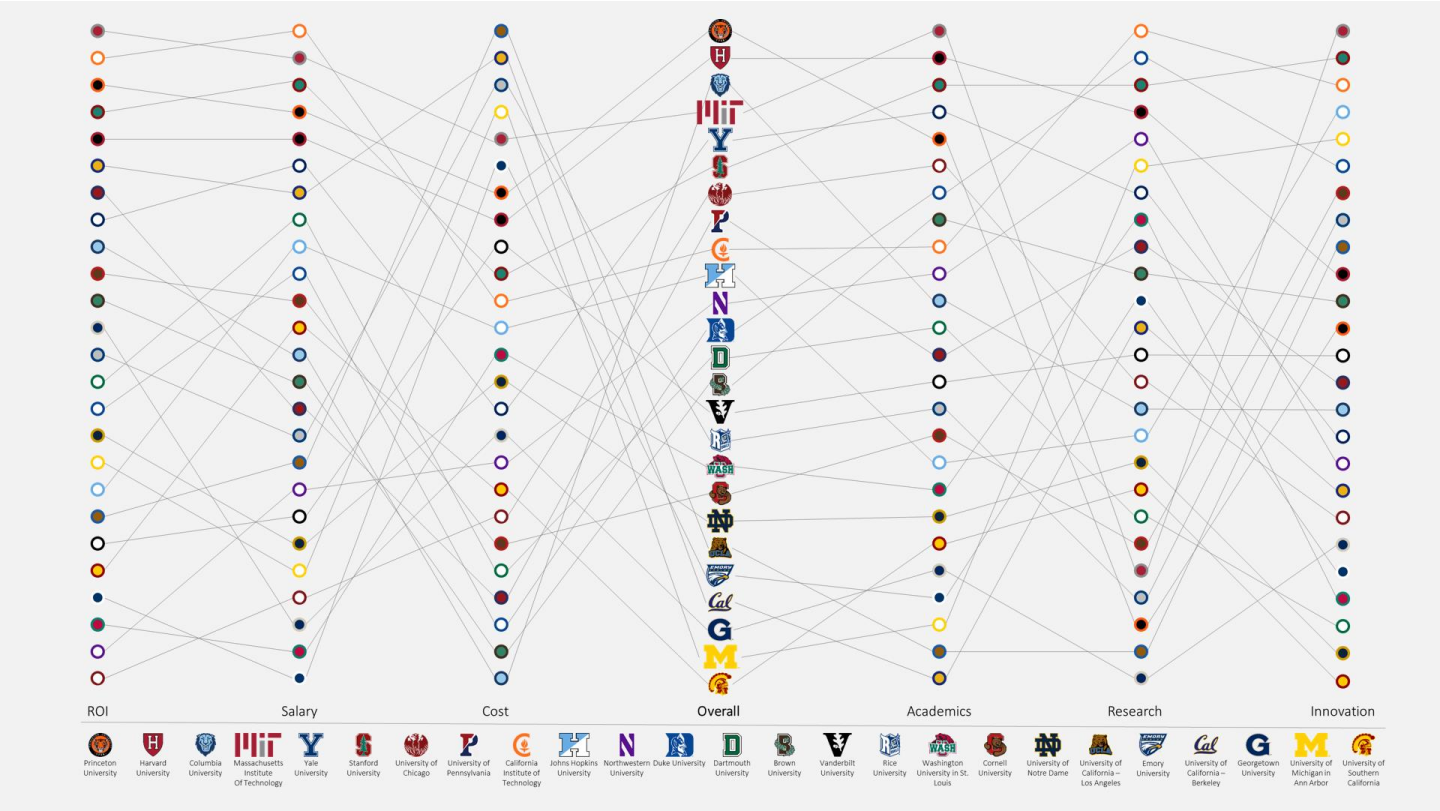


Figure 3. Final visualization of top 25 colleges and universities ranking across 7 domains: Overall, Academics, Research, Innovation, Cost, Salary, and ROI.

Discussion

The undergraduate university research process can be time consuming and labor intensive and is exacerbated by the fact that university information is often fragmented across sources and even fragmented within those sources across different webpage, lists, and league tables. Additionally, these lists and league tables only show limited information for a small number of universities on the screen at one time. This display/information fragmentation and the keyhole effect increase the cognitive load of the research process by forcing users to hold information in memory and perform mental calculations or comparisons, or externalize that information (by aggregating rankings in a separate spreadsheet, for example), further increasing the labor needed to complete the task.

A bump chart, a visualization often used to show change in ranking over time, can be used to show rankings of a larger number of universities across various domains and enable easy side-by-side comparison. Placement of the university icons in the ordered list allow salient characteristics such as rank and standing to emerge. The slopes of the lines allow salient features such as amount of change in ranking to emerge and become apparent to the user, especially between domains, as well as potential overall correlation between domains.

Limitations

There are a few limitations to the analysis performed and the created visualization. First is that the data used to create the domains were collected from various sources. We are not privy to the methods or data used to create the ranking lists among the different sources, and so we cannot really determine whether the domains themselves are comparable. This is indicated by the relatively low amount of correlation between the different domains. It is possible that if the domains were re-created from raw data, we would see more discernable patterns in the rankings between the universities across the different domains.

Additionally, while the bump chart is advantageous for comparing multiple entities, it certainly has its limits. Originally, the visualization was created using a sample of 45 universities; however, this visualization quickly became unwieldy and the amount of entities hampered the visualization's readability. It would be advantageous to explore other ranking visualizations that allow a larger sample to be represented.

Collaboration on Scientific Papers in Assistive Technology as Compared to All Papers

Erin Higgins
Contact: elh108@pitt.edu
The University of Pittsburgh

Figure 1. Collaboration on All Available Scientific Papers. The Microsoft Academic Graph Dataset [1] was used to generate this visualization. The Microsoft Academic Graph (MAG) is a heterogeneous graph containing scientific publication records, citation relationships between those publications, as well as authors, institutions, journals, conferences, and fields of study. This first visualization used every paper within the dataset that included latitude and longitude information and listed multiple uniquely located collaborators for the paper. Any paper that was created entirely by one university or city was not included. Quite a few papers were also missing latitude and longitude information and were also not included. This led to a dataset of just over 1 million papers. The affiliations table was used to gather the latitude and longitude of the various contributors. Python's greatest circle calculation from the mpl_toolkits was used to create the connections. Lighter pink correlates to more collaboration between the two locations. This visualization does not show collaboration within the universities, but simply their external collaborators.

Figure 2. Collaboration on All Scientific Papers with "Assistive Technology" as the subject. The second map visualization used the field of study table to only includes papers in which assistive technology was listed as the main field of study. This led to a dataset with 10,936 papers. The exact same Python tools were used to generate this map simply replacing the larger full paper dataset with the smaller AT specific dataset.

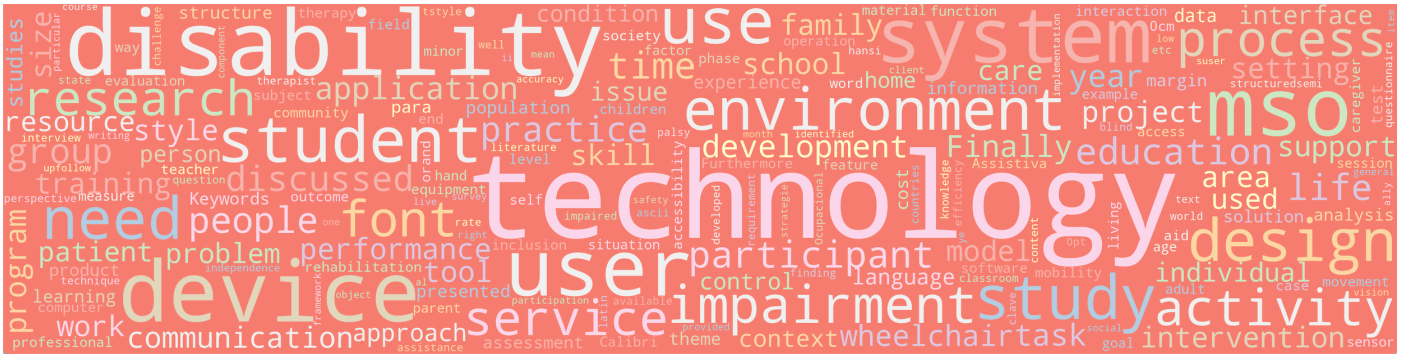


Figure 3. WordCloud of the Abstracts of All Papers with "Assistive Technology" as the Subject. The third figure is a WordCloud of the most used words in the abstracts of the papers related to assistive technology, not the full dataset. This was generated by pulling the inverted index abstracts from the appropriate tables in the MAG. The Python WordCloud package was used to generate it once the data was cleaned appropriately. Some words, such as "background", "methods", and "results" were removed, as they showed up in every paper but were not relevant to the context.

Significance

Assistive technologies (ATs) are needed by an estimated one billion individuals worldwide to participate fully in society and live active, independent lives; without them, individuals are often excluded from society, do not have access to basic opportunities such as education and jobs, and are at a higher risk of being poor and unhealthy. This number is expected to increase to 2 billion by 2030 [2]. ATs include any device that allows an individual to participate in all aspects of society. Some classic examples are wheelchairs or screen readers. This research examines this need and helps to explore why certain countries, such as many countries in Asia, have a much harder time accommodating the needs of AT users. By analyzing over 1 million papers and comparing connections between researchers, we can see where possible hubs of AT innovation are located and where this is not happening. It is very clear that the US and Europe do considerably more work in the AT space than other parts of the world. Unfortunately, it seems as if these areas are also more likely to collaborate among themselves than with other countries, which also may be contributing to the lack of access to AT in those parts of the world. The abstracts, analyzed with a WordCloud, show that AT specific abstracts deal with concepts that are universal. This means that if more individuals innovating would consider all users when designing, there might be less of a gap of access and community involvement around the world.

Findings

1. *Many of the academic hubs of the world are not doing work in AT.* In the first figure, it is obvious that Asia is a hub for academic collaboration. The second figure, however, tells a different story. These countries are far less involved in AT than Europe or the US. Countries in Africa and South America also drop in number of collaborations considerably when looking at AT. Unfortunately, in these parts of the world, individuals with disabilities are less encouraged to participate in society. Sometimes they are even considered shameful and encouraged to remain at home (3). This is directly reflected in the amount of publication collaboration that occurs in these locations.

2. *The United States and Europe collaborate more within themselves than other countries overall and in AT.* As seen in the first and second figures above, Europe and the United States are much more interconnected than other countries. The connections are lighter, meaning that they collaborate more often and then are so dense that you can visualize the edges of the countries. They collaborate around the world, as well, but there are far less dense connections. The dense connections within countries is less helpful for new researchers, as it is hard to discern where they should be looking for collaborators.

3. *Asia and Australia collaborate around the world more than within themselves when it comes to AT.* The opposite is true in Asia and Australia. These two areas have many more connections going out of the countries than within themselves in figure 2. Japan is a bit of an exception, but they still have a great number of connections going out as compared to within. This has the potential to be more helpful for new researchers in these areas so they can see where there are other people working and attempt to begin collaborating within their country.

4. *South America and Africa have "hubs" whereas other countries have more spread-out collaboration.* In South America and Africa, it is more obvious that there are small areas, such as South Africa and Brazil, where most of the collaboration is taking place in figures 1 and 2. These places have been partners on many papers externally, but not very many collaborations have occurred within the countries themselves or in other countries on the continent. These continents have many more hubs for scientific research in figure 1 than in figure 2 leading to an assumption that they are also not doing as much work in AT as they are in other topic areas.

5. *The topics from the AT abstracts are not exclusive to AT.* The topics seen in the WordCloud show some items that every person should be concerned with. Topics such as "technology" and "user" come up frequently. Too often, research is concerned with innovation and not thinking of their user. If user centered design became more mainstream, perhaps all technology could become assistive technology.

6. *The abstracts are very focused on user centered design.* From figure 3, we can see that the abstracts are focused on users. Students and users are two of the biggest categories in the figure. This is a difference from other disciplines. For example, physics would often be far less user focused than this category. It shows that to be published in the field, it is important to consider the individuals using the product you are creating.

7. *Wheelchairs are mentioned most often in terms of devices.* Computer interfaces and other wireless technologies are listed in the WordCloud, but wheelchair is the largest device. This means that of most of the papers published, wheelchairs are the device that is most discussed. This will be helpful to individuals who are attempting to distinguish themselves within the field from choosing a topic that is already so researched.

Data

The Microsoft Academic Graph (MAG). MAG is a heterogeneous graph containing scientific publication records, citation relationships between those publications, as well as authors, institutions, journals, conferences, and fields of study. It is freely available for download and storage in a Microsoft Azure Student account. This was the only sources of data for this research.

Method

Many steps were followed to generate the data for these figures. First, an account on Microsoft Azure was created to store the MAG data. After this was accessed, a Databricks account was created and attached to a cluster on Microsoft Azure. From there, the cluster was used to write SQL queries to find specific data.

Figure 1 Generation: All Paper Connections

First, a query was run to download all paper IDs. The PaperAuthorAffiliations table was used because it did not contain all of the extra details that the Papers table contained and took less time to generate. From there, this table was joined using an inner join on the AffiliationId field to

the Affiliations table. This returned a table containing the Latitude, Longitude, and PaperId for every paper in the dataset. If latitude and longitude did not exist, the paper was ignored. From this table, PaperId was used to match papers with different latitudes and longitudes. If there was more than one set of latitude and longitude for a single PaperId, this was considered a collaboration and continued as a part of the dataset. If there was only one instance of the PaperId, it was not included in the dataset. After pruning like this, the dataset contained just over 1 million individual papers to mark on the map. The data was exported as a CSV.

In order to generate the map figure, Python was used. The following imports were required to generate the graph: csv, math, numpy, pandas, mpl_toolkits, pyplot and matplolib. To generate the map, the CSV file, generated as described above, was parsed and used to create a python dictionary that was a list of paper IDs along with a list of tuples containing the latitudes and longitudes that contributed to the publishing of the paper. From there, a Basemap was created from the matplotlib toolkit. The Miller Cylindrical Projection was used to plot the map. Country and continent outlines were then generated on this map. From there, the dictionary of latitudes and longitudes was iterated through and the Python "DrawGreatCircle" function was used to map the projections. This function will calculate the shortest possible arcing path between two points. and is used as follows:

```
line, = m.drawgreatcircle(
      float(latlong[1]), float(latlong[0]), float(latlong[3]),\
      float(latlong[2]), linewidth=0.5, color=color)
```

An important next step was to calculate cut points for certain paths. Because the Earth is a globe, some paths were wrapping around the map. To ensure this did not print, any path whose vertices were shown to be over a certain length were split and the second half appeared on the other side of the map instead of cutting off half way through the line as shown below:

```
cut_point, = np.where(np.abs(np.diff(path.vertices[:, 0])) > 30000e3)
                if len(cut_point) > 0:
                    cut_point = cut_point[0]
                    vertices = np.concatenate(
                              [path.vertices[:cut_point, :],
                              [[np.nan,np.nan]],
                              path.vertices[cut_point+1:,:]])
                    path.codes = None
                    path.vertices = vertices
```

The image was then printed and saved as a .png file.

Figure 2 Generation: AT Specific Paper Connections

A very similar method was used to generate figure 2. The only difference was the generation of the data to be imported. After papers were downloaded, a second query was run to select only the papers whose subject matter was "assistive technology". This was done using the FieldsOfStudy table. A query was run on that table by selecting only the FieldOfStudyId of papers whose NormalizedName was equal to "assistive technology". From there, that FieldOfStudyId was joined with the PaperFieldsOfStudy table to generate a list of all PaperIds where the FieldOfStudyId matched the Id for "assistive technology". From there, the exact same method

was used for generating the latitudes and longitudes, which was exported as a CSV, as well as the same code to generate the map.

Figure 3 Generation: AT Specific WordCloud Generation

This figure required different methods. To generate the abstracts, the table of PaperIds specifically related to AT from the previous segment was used. These Ids, instead of being passed into the Affiliation table, were passed into the PaperAbstractsInvertedIndex table. This generated a list of abstracts as a CSV that were from all papers with "assistive technology" as their subject.

From there, Python was used again to generate the WordCloud. The following Python tools were imported for the task: csv, json, matplotlib, and wordcloud. This script opened the CSV generated as described above. From there, it looked at each row and created a text string of each word in the abstract. The Python tool STOPWORDS was used to ensure that frequently occurring and insignificant words were not included in the analysis. Stemming was also used to ensure that each version of the word was counted in the same category. From there, the Python tool WordCloud was used by inputting the string created from the abstracts and generating the WordCloud. In this section, a StreamGraph was attempted but the data required too much cleaning and no easily understood StreamGraph could be generated.

Discussion

Though it is not discussed as often as most other topics, assistive technology innovation is a growing problem for the world. As people live longer, more and more individuals eventually will need AT to continue functioning in society. Money and time must be put into the innovation of these products or the world will face a crisis in which millions of individuals are unable to get the products that they need to live and continue as a part of society.

This research shows a very important and unfortunate truth: not much collaboration is happening in the world of AT. Outside of the United States and Europe, there seem to be single institutions in other countries that are attempting this sort of research. These small hubs in different parts of the world, however, do nowhere near as much internal or external collaboration as compared to general publications.

This research reveals another important truth: assistive technology seeks to solve the same problems as other designs. The topics seen in the WordCloud such as "technology" or "student" are universal to most individual's life experience. Unfortunately, most engineers are not taught to consider all users when innovating. If all design programs taught the importance of accessibility and user centered design then perhaps the world could become a more inclusive place. Places such as China, who are producing a great amount of outstanding work in many technologically advanced areas, as seen in figure 1, are producing nearly no innovation in AT, as seen in figure 2.

Some of this might have to do with the fact that the United States and Europe have specific funding organizations that are focused on AT. In the US, NIDILRR is a federal organization that

funds innovation in AT (4). Unfortunately, the market for AT is very small and fragmented. This means that innovating in these areas will not generate a great revenue for anyone. Because of this, countries where innovation is funded without expectation of profit, like the US, should take the lead in spreading this information and collaborating more throughout the world. Some countries, however, will most likely not see this sort of financial support in the near future.

In many countries around the world, people with disabilities are shut away from society. Often this comes from outdated ideas about disabilities such as a perception that they are related to misconduct in a previous life. Because of this, in some instances, a family member with a disability is perceived as a disgrace to the family (3). Because of this lack of collaboration seen very obviously in this research, these ideas are not changing as quickly as they could be. This research shows that it is time for the countries who are doing research in these areas to seek out collaborations to encourage a change in attitude around the world. People with disabilities are struggling to be a part of society in all parts of the world, but especially in the places where we can see no collaboration happening in figure 2. If individuals who are doing this research could seek out collaborations in the parts of the map where it's not happening, perhaps the world could be made more inclusive for all people.

References

[1] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang (2015). An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of the 24th International Conference on World Wide Web* (WWW '15 Companion). ACM, New York, NY, USA, 243-246. DOI=http://dx.doi.org/10.1145/2740908.2742839

[2] Assistive Technology (2018). Retrieved from: https://www.who.int/news-room/fact-sheets/detail/assistive-technology Accessed February 12, 2020.

[3] Parker KJ (2001). *Changing Attitudes Towards Persons with Disabilities in Asia.* Disability Studies Quarterly. 21,4;105-113.

[4] About the National Institude on Fisability, Independent Living, and Rehabilitation Research (NIDILRR) (2020). Retrieved from: https://acl.gov/about-acl/about-national-institude-disability-independent-living-and-rehabilitation-research

American Conference Bibliometrics

Jingyi Sun
JIS111@pitt.edu
The University of Pittsburgh

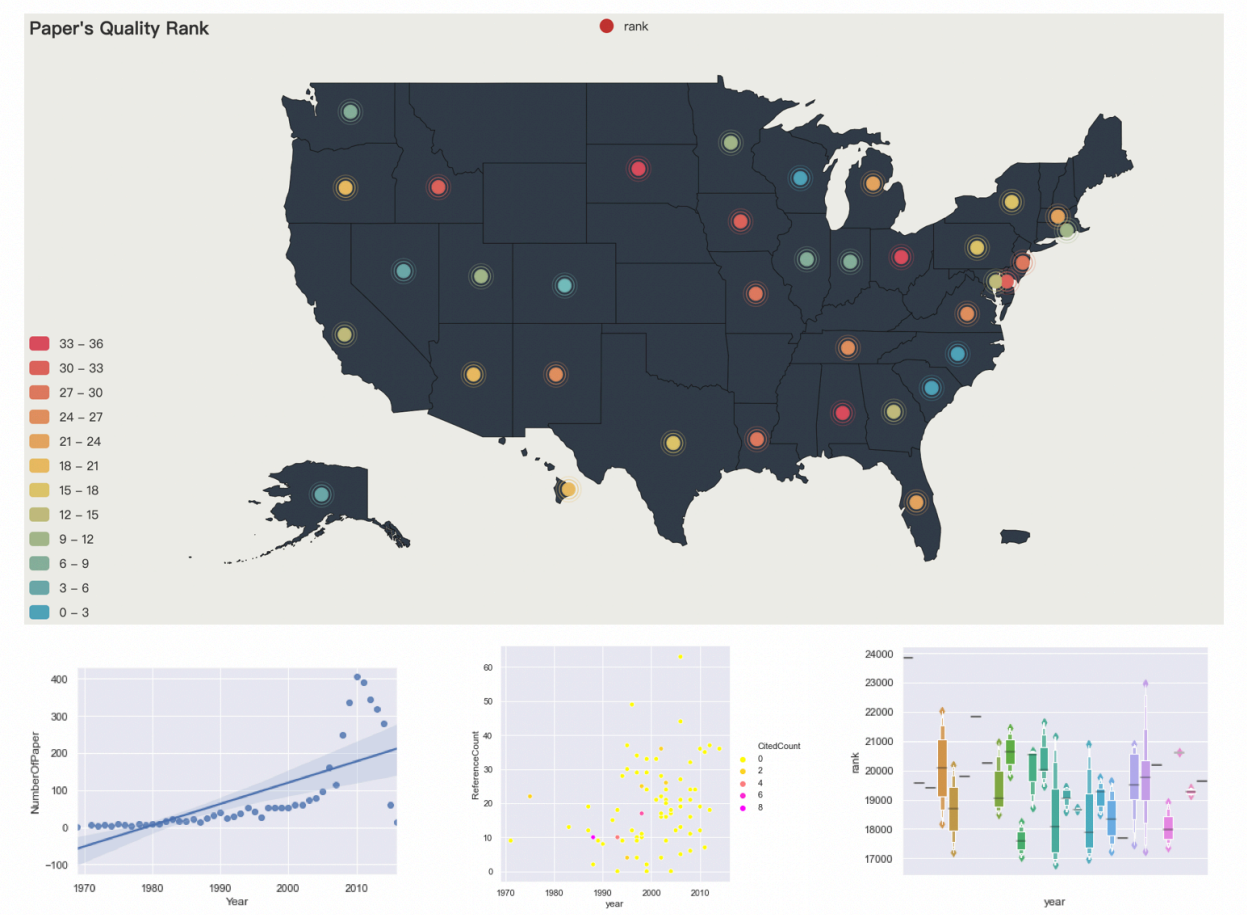


Figure 1. American Conference Bibliometrics. Three text files are used (see Data for more information). The main text file is paper.txt, which contains 1048576 papers' information published all over the world from 1801 to 2016. In this project, we mainly use the data of papers publish in American conference. So, we merge the text file of conference instance and papers, then selected 15894 papers which meet our requirement. As we believed that the higher the ranking, the better the quality of papers, we do the visualization of the average ranking of the papers according to the state they published in to see the geographic distribution of the quality of American conference papers. Locations are colored by the average rank of conference papers. **b,** a picture of the relationship between time and the number of papers to mining the rule of paper growth. x axis representation the number of year while y axis represents the number of papers published each year. Papers number ranged from 2 to 403. **c,**

the number of citations of a paper is an important factor in evaluating the quality of a paper. We combine the conference instance with paper reference data to establish the relationship between the paper and its cited papers. Group the data and calculate the number of paper id they were cited. Using the data as the number of citations. Then draw the relationship between the number of citations and the number of references. To help us better evaluated the quality of papers. **d,** a boxed plot conveys information about central 50% and extent of paper's rank information each year. Except the quartiles, this plot offers more precise estimates of quantiles beyond the quartiles, which is show the detail information about the tails.

Significance

This research project aims to have an overall grasp of the publishing status and quality of American conference journals and provide some references for the development of knowledge organizations and academic conferences. Motivated by the statistical methods of bibliometrics. We mainly analyze the data from time distribution, location, ranking, number of citations and the number of references to discovery the development and trend of American conference papers after 1997.

Findings

1. Compare to the west, the conference journals published in the east are of higher quality but not stable. The top three rank states are all the east. The average rank of top 1 Wisconsin is 20420, top 2 North Carolina is 20603, and top 3 South Carolina is 20658. However east also have the last two. This situation may be caused by various factor. Generally, we assumed that the quality of paper usually tends to be higher in the place with better academic atmosphere or top university. In the west, like California where have many top universities, it has published the most conference papers with 10254 in those years and the average rank of those papers are relatively high.

2. From 1969 to 2010, the number of conference journals published in the United States increased exponentially, from 2 papers per year to 407 papers per year. Then the number remained relatively stable for the next three years and drooped sharply.

3. The most papers are only cited by one other paper that is also published in American. The number of citations of a paper often represents its quality. However, in the United States, the phenomenon of mutual citation of papers is not particularly obvious.

4. The quality distribution of papers is not necessarily related to the published number and the time of its publication. In 2010, totally 403 conference paper have been published. Whereas its rank ranged from 17000 to 23000. It proves that the paper review standards in the United States have not changed year by year.

Data

The Microsoft Academic Graph is a heterogeneous graph which store over 120 million basic information and connection between scientific publication records. It is the largest publicly available dataset, and also contains citation relationships between publications, and institutions, journals, authors, conferences, and fields of study. This graph can be received on Azure storage account at https://docs.microsoft.com/en-us/academic-services/graph/get-started-setup-azure-data-lake-analytics. I used existed data in version 2019-03-22. It can be directly download through zenodo at https://zenodo.org/record/2628216#.X8MXspMzaRs.

Three main graphs were used in my analysis. They are papers.txt, conferenceinstances.txt, and paperreference.txt.

Method

Bibliometrics

Bibliometrics is a branch of library and information science that uses mathematical and statistical methods to describe, evaluate and predict the status and development trend of science and technology with the help of various information of papers. Since the 1920s, mathematical statistical methods have been introduced into paper analysis. With the development of computer technology, data science has been applied to bibliometrics since the 1950s, made its theory and application have been greatly developed.

There should be some regularity behind the growth of scientific paper. Price, D.J. de Solla (1976) defines the exponential growth of the paper. The existence of the growth pattern is obviously universal and long-term. $F(t) = k/(1 + ae^{-bt})$, ($k>0$, $a>0$, $b>0$). $F(t)$ is the function of time, k is the cumulative amount of documents when t tends to infinity, a is conditional constant, and b is time constant. From the perspective of mathematical analysis and statistical examples, the exponential function is exactly a mathematical expression of the number of scientific papers increasing with the pass of time, which is consistent with the statistical results of scientific paper in a certain historical period. However, this method cannot be used to predict the increase of the number of papers. According to the rule of the exponential function, the increase in scientific documents will tend to be infinite over time. Obviously, it is difficult for human beings to do the research to meet the requirement of unlimited growth of papers.

Discussion

When using bibliometrics to analyze the development process of the conference paper, according to the growth or decay rule of the paper, the annual statistical analysis of the number of related papers can reveal the current development status of the science. In the past 30 years, from 1970 to 2010, the number of conference papers has increased exponentially.

During this period, there were three obvious decline points in the number of papers issued, namely 1991, 1996 and 2007. The rest of the time remains in a state of growth. However, after three years of relative stability, the volume of publications in conference paper has drop significantly. This situation does not conform to the rules of the increase in the number of papers and the development of science and technology. The specific reasons need to be further analyzed.

In terms of the geographical distribution of the number of conference papers published, the top states include California, Florida, Miami, Washington, etc. It shows that these states hold more conferences. Especially California, as there has the large number of top universities and better environment, the number of conference articles issued is far ahead of other states. But the ranking of the conference paper is another situation. The top 3 ranking state are Wisconsin, North Carolina and South Carolina, which is totally difference situation according to the number of papers. From this we can come up with that perhaps the academic quality of the conference in the East of the United States is higher, although there may be some objective reasons that there is no way to publish too many conference articles. Or it may be that the review criteria for conference papers vary with the state. Some institutions can also review the criteria based on this phenomenon.

It can be seen that overall academic environment in the United States is getting better over time. The changes in the paper ranking are slowly getting smaller and the overall rankings are getting higher. Citation number should be an important factor account for the quality of the paper. However, In the analysis of American conference paper, the number of citations still have some limitations, which cannot reflect the actual citations. Further study needs to be done.

So Eun Lee
INFSCI 2415
sol47@pitt.edu
University of Pittsburgh

Leader or Collaborator: Assessing Individualism in the Scientific Community

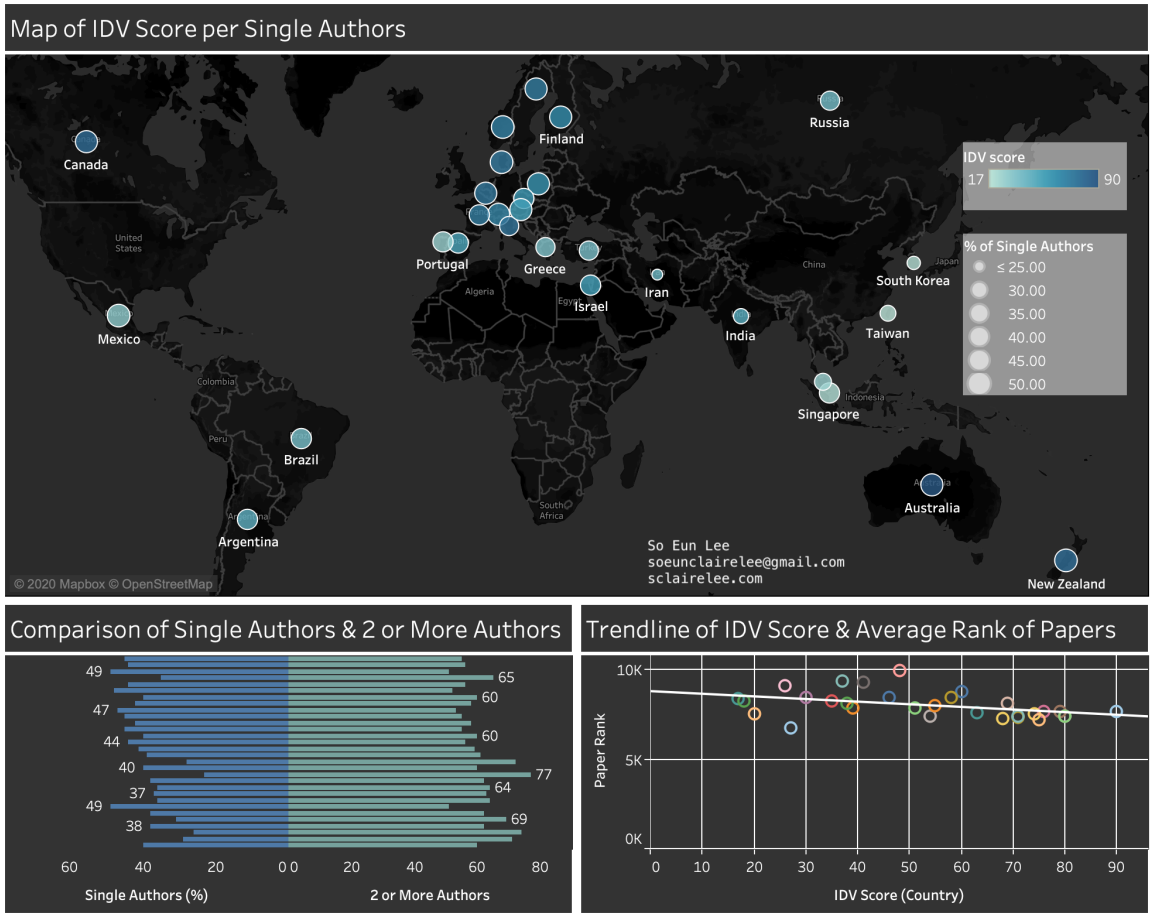


FIGURE 1. Leader or Collaborator. Two main datasets have been used to generate these visualizations. One dataset provides the geographical identifier for where the papers are being produced, which is linked to another dataset that contains information about the papers and authors through the geographical identifier. We focus on the top 30 countries that produce the most amount of scientific papers. The data for each country is extracted to count the number of times a distinct author sequence number appears in the dataset, which also represents the number of authors who contributed to a paper. Approximately 81% of the papers have fewer than five authors, so we focus our analysis on up to five authors. **A.** Each country is assigned an individualism score based on a Hofstede scale – the higher the number on the scale and the dark the shade of the dot, the more individualistic the country is. The size of the dot indicates the percentage of authors who published their scientific papers as a single author. The two sets of information can be collectively used to observe how individualism affects the likelihood of individualistic tendencies in producing scientific articles. Though not included in the

visualization, the results show a positive correlation of 0.60 (r-squared: 0.36) with a p-value < 0.001 between the individualism score and the percentage of papers with single authors. This shows that the more individualistic a country is, the more likely it is to have papers that have a single contributor. **B.** The dual axis chart on the bottom left visualizes the percentage of papers within a country that have a single author vs. two to five authors. The countries are placed in a descending order of the number of papers produced. We observe that the increase in the number of authors does not lead to a greater production of papers. **C.** We display the statistical relationship between individualism scores and the average rank of papers produced in each country. The results show a negative correlation of -0.42 (r-squared: -0.17) with a p-value < 0.001. As the number of authors increase, the average rank of the papers decrease.

SIGNIFICANCE

Collectivism is defined as "the degree to which individuals are integrated into groups" [1]. A member in a collectivist society is taught to emphasize the "we" and consider themselves as a part of a larger community (emphasis on collaborators), whereas a member in an individualist society focuses on the importance of self and placing personal tasks over relationship (emphasis on individual leaders) [1]. This study visualizes the way the scientific community is affected by such cultural communication. By analyzing the number of contributors to the scientific papers from some of the countries that produce the greatest quantity of papers, we demonstrate that a country that is rated as more "individualist" on a Hofstede scale is more likely to produce papers with a single author from their own country, while collectivist countries have a greater percentage of collaborators working on a publication. This result demonstrates that the amount of collaboration happening in the scientific field can be predicted by the measure of individualism in a country. In addition, by demonstrating that the increase in the number of authors does not necessarily lead to an increase in the quality of papers (average rank) or the number of papers, we show that more research collaboration does not necessarily yield more productivity and success despite common belief.

FINDINGS

The more individualistic a country is, the more likely it is to have single authors producing a scientific paper. The Pearson correlation coefficient between the individualism score and the percentage of single authors (r-squared) is 0.36 (p-value < 0.001).

The higher the individualism score, the higher the overall rank of the country's papers. The Pearson correlation coefficient between the individualism score and rank of papers (r-squared) is -0.17 (p-value < 0.001).

DATA

The datasets are generated by the Microsoft Academic Graph. All datasets used for the purpose of this study can be found at: https://zenodo.org/record/2628216#.X8PCX6pKiv6

Affiliations dataset. This dataset contains scientific publication records for 192 countries, including the geographical identifier (GridId) of the affiliations that produced the scientific

articles. This dataset also provides the rank of each paper, where rank is measured using its relationship to other articles (e.g., being cited in another article in the dataset).

GridID dataset. This dataset was downloaded from https://www.grid.ac/downloads and was joined with the Affiliations dataset to connect the actual geographical coordinates of the research institutes so that the count of articles could be associated with the correct country.

Paper-Author-Affiliations dataset. This dataset contains information regarding the authors and provides the identifier for each paper in the records. It was joined to the Affiliations dataset via AffiliationId. The number of authors that contributed to a specific paper was determined by counting the distinct number of times the id of the paper appeared in this dataset.

METHOD

Geert Hofstede is a pioneering researcher in cross-cultural psychology, and a large focus of his research has been on studying collectivism and individualism in different countries. We obtained the individualism score for each country from a public dataset found here: https://www.hofstede-insights.com/product/compare-countries/. The score is given from a range of 1 to 100; the higher the score, the more individualistic the country is.

Due to the sheer size of the dataset, we chose to focus on the top producers of scientific papers. We picked a sample size of n=30 to ensure the results of the study are meaningful. Extreme outliers, such as U.S. and China, were excluded from analyses.

Each scientific paper was connected to the country where it was produced via GridId. The number of times each paper appeared in the dataset was counted to represent the number of authors who contributed to the article (i.e., each time a distinct individual contributes to a paper, this information is recorded in the dataset). About 80% of the publications had an author sequence number (the order in which their names are listed on paper) ranging from 1 to 5, so we chose to focus our study on up to 5 authors. The data for each country was extracted and downloaded into files, then we used R to generate 30 new datasets with paper identification and the number of authors who contributed to them. We joined the individualism score for each of the countries in the dataset along with the average rank of the papers produced in the country.

DISCUSSION

This study shows the difference in cultures that shape research collaboration. The cultural norms of a society undoubtedly bleed into various aspects of life. One of the most blatant examples of this may be in the professional sphere, and the collectivist culture's affinity toward collaborative work is likely to occur in the research setting, as well. However, this focus on working together may come at a cost to performance.

There is a long-standing belief that research collaboration has a positive effect on the productivity of publications [2]. However, the results of this study show that collaboration doesn't seem to have a positive effect on the quality of paper; in fact, the findings seem to

indicate otherwise. This supports the findings of another study that demonstrate that the number of collaborators is not a "significant predictor of publishing productivity" [2]. When there are multiple authors who are involved in the publication efforts, researchers may also be introducing various factors that impede their work (e.g., increased need for communication, making compromises to accommodate diverging interests, etc.). Further research will need to investigate the factors that detract from the performance of collaborative work.

One of the limitations of this study is that we solely focus on the individualism score to measure a culture's tendency toward individual/group work. While this score may be a predictor, there may be other factors in a society (e.g., the amount of funding) that discourage or encourage collaboration.

There is an African proverb that states, "If you want to go fast, go alone. If you want to go far, go together." As wise as this saying seems, in the world of scientific papers, this may not be the case.

Reference List

[1] G. Hofstede, “Dimensionalizing cultures: The Hofstede model in context,” *Online Readings in Psychology and Culture*, vol. 2, December 2011.

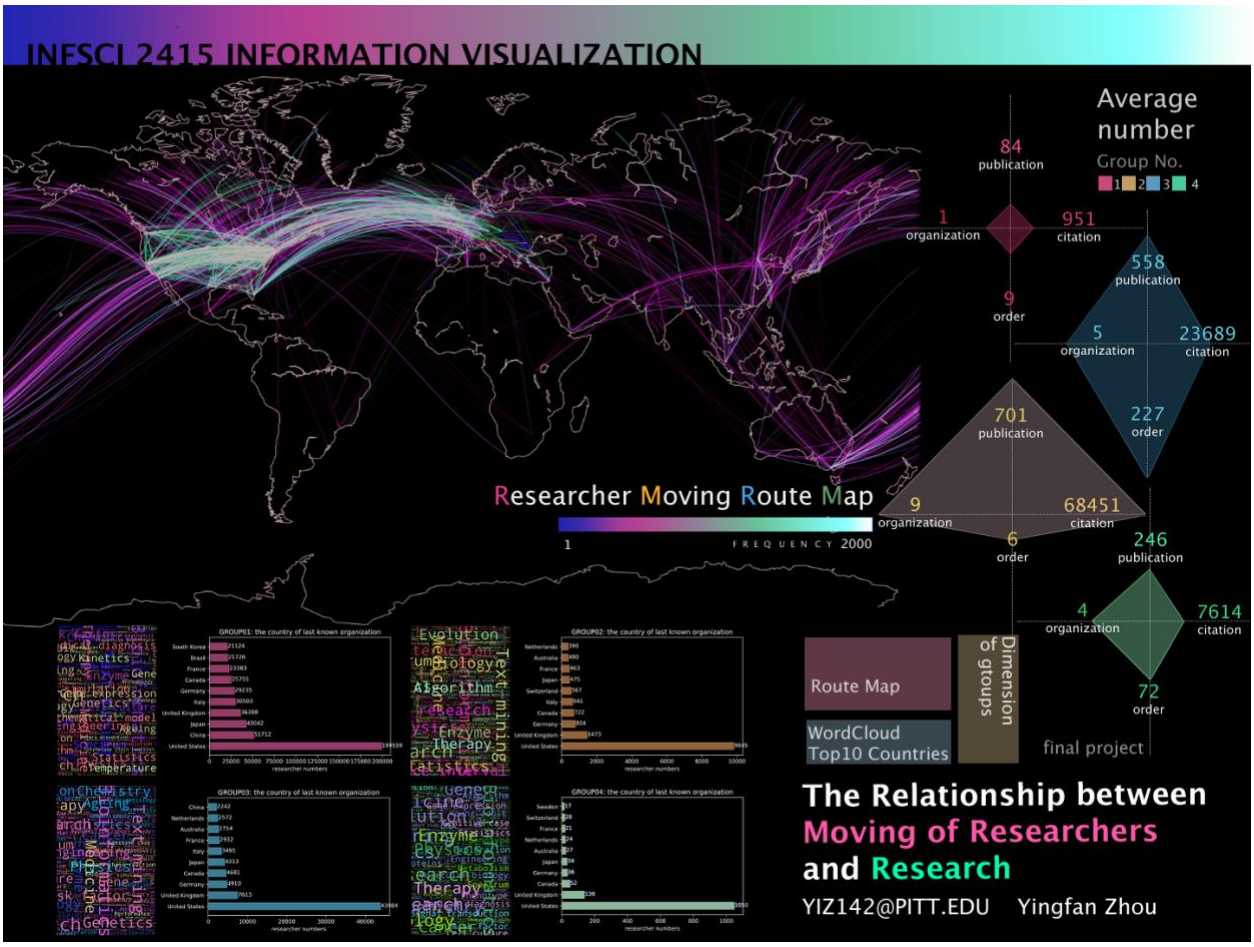
[2] S. Lee, B. Bozeman,” The Impact of Research Collaboration on Scientific Productivity,” *Social Studies of Science*, vol. 35, pp.673-702, October 2005.

[3] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B. Hsu, and K. Wang. 2015. “An overview of Microsoft Academic Service (MA) and applications.” *Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion)*. *ACM, New York, NY, USA*, pp. 243-246, doi: 10.1145/2740908.2742839

[4] K. Wang et al., “A review of Microsoft Academic Services for science of science studies”, *Frontiers in Big Data*, 2019, doi: 10.3389/FDATA.2019.00045

How do the researchers move to different affiliations?

Yingfan Zhou
YIZ142@pitt.edu



The MAG datasets which contain the information of papers, authors, organizations are used to generate this graph. The graph describes the researchers who published more than 30 papers from 2008 to 2017 in any of research areas.

Researcher Moving Route Map

This graph draws the route of researchers moving. Frequent moving happens inside between US and Europe. We can observe the considerable moving happens between China, Australia, Japan, Europe and US. Many researchers will change their organization inside US.

Dimension of groups

This radar chart describes the features of authors in four groups, including number of publications, number of citations, number of author order in the paper, and number of organizations stayed. Group 1 contains 803532 authors who almost do not change their locations. 99163 authors in Group 2 frequent change their locations, published a lot of papers which are largely cited. Group 3 contains 18597 authors whose author order in the paper is big. Group 4 contains 1562 authors whose features are in the median of all groups.

WordCloud & Top10 Countries of last known organizations

Wordcloud graph describes the frequency of keywords of papers in 4 groups. I found that most of words are related to computer science and medicine. But the research area are different in 4 groups. In order to know how researchers are distributed, I drew the bar chat to show the number of countries contains last known organizations. We can find that US is top1 in any of groups. And the sequences of counties in different groups are different. Most contains US, Europe, China, Japan, Australia, which can be also identified with moving route maps.

Significance

This project responds to the relationship between moving of researchers and research. The results demonstrated that the number of organization where researchers stayed is positively correlated to the number of publications. Frequent moving between different organizations may increase the productivity of research.

Findings

1. *Most moving happens between two different countries*.
2. *Most of researcher will choose moving to the organizations in US and Europe.*
3. *The number of organization where researchers stayed is positively correlated to the number of publications.*

Data

Microsoft Academic Graph dataset on Azure (https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/) includes information about papers, authors, affiliations and so on. In this project, I defined the active researches as the authors who published

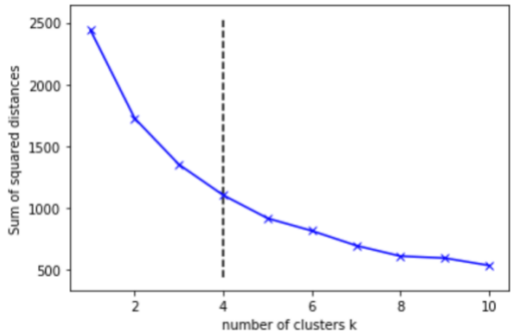
more than 10 papers from 2008 to 2017 in all research areas. There are 3451993 authors who fits the definition. Considering the large data scale, I calculated the median and quartile of numbers of paper published by each researcher and chose the upper quartile to filter out the authors who actively published papers in recent 10 years. The final dataset I used in this project contains 922855 researchers who published more than 30 papers from 2008 to 2017. With the data of papers, I listed the affiliation where author published papers in sequence to study the moving routines of researchers.

Method

Cluster analysis

I did the K-means clustering on the authors to clarify different groups of researchers based on the number of publications, number of citations, the order of author in the paper. Since the distribution of order number of authors in papers is skewed, I used the median number of order numbers in all papers written by each researcher.

The K value is determined by the result of elbow method which is shown in the below graph. When K=4, the change of curve become flatter. So, I choose K=4 as group numbers.



Then, I calculated the average number of publications, citations, the affiliations where authors stayed, the order of author in the paper to identify the difference between groups. Because the data is large and the difference between average number and median number is not different. Additionally, I counted the frequency of research interests in different groups and the last affiliation we knew about the author.

Discussion

In this project, based on the graph and result, I find that the change of organizations may be positively correlated to the productivity of research which is measured by the number of publications. But I do not get the relationship between the paper's quality which is measured by the number of citations and moving movements.

Here is one sampling problem. Because I sample the researchers who published more than 30 papers in 10 years, it narrow research area. Most of research area includes Medicine and computer science, it is hard to find the difference of research areas in different groups. And, it may lead to a bias on the conclusion of moving movements.