# Data Noobs

*Abhijit Mahapatra, Rohit Dube, Sambandh Dhal, Swarnabha Roy, Tushar Pandey*

## Executive Summary

Website: https://share.streamlit.io/pandey-tushar/tamids-22/main

This project focuses on analyzing the impact of collaboration on the research outcome of Texas A&M University. It is a data-driven research that tries to illuminate the patterns of collaboration across disciplines within the university as well as the relationship with other universities. We try to estimate the influence of certain buzzwords in research that can help the university fetch more research grants.

The main takeaway of the project can be stated as follows:

1. The cumulative altmetric score was used as the criteria of effective and impactful interdisciplinary research as it was an indicator of high multidisciplinary research collaboration, highest number of research papers and highest media attention. In order to restate this claim, the data from 133 departments were used for analysis and it was proven that the Department of Electrical and Computer Engineering which had the highest cumulative altmetric score had the highest number of standalone publications as well as the highest number of publications in collaboration with other departments.

2. For each Department in Texas A&M University, we tried to analyze the grant data as a visualization graph for every decade starting from 2000. The length of the funding has also been stated in the bubble graph which is demonstrated by the size of the bubble in the analysis.

3. Another metric devised by the team was coined as “Impact score” which is associated with every “buzz word” devised from the publication data containing the abstract and the grant data containing the details of the funding agency which provided the grants. Multiple “buzz words” were computed per department and for every buzz word, the impact score was calculated taking the normalized value of funding associated with each word, number of total buzz word citations, the number of publications associated with each buzz word and the total number of citations received by the Department. This has been explained in detail in the later part of the report.

## Problem Statement

The main problem which has been identified as the motivation is the channelizing of grants and collaborations among the departments within Texas A&M University resulting in impactful publications. The other insight which we seek to address is the fact that future collaborations can be mustered based on the past data so as to boost the overall ranking of the university by acquiring meaningful collaborations within the University and with other participative universities for acquisition of funds.

# Data Collection and Preprocessing

The corpus used to design our approach was combined using different datasets extracted from

a. scopus (Github link),
b. dimensions.ai (Github link),
c. scholars TAMU (Github link) and
d. altmetric.org (Github link)

# Data Exploration

Before diving deep into the analysis, we tried to estimate the impact of research using four main-terms per abstract word that have been defined per department.

- Publication score:
- Citation score
- Normalized funding
- Impact score

Impact Score = Normalized(Funding(Topic for dept)) x [no of publication (for topic in dept) x Total citation for topic]/Total citation for dept

$$\text{Impact Score } (\mathscr{IS}) = \frac{\widehat{F}(Topic) * ||P(Topic)||_0 \; * \; ||C(Topic)||_1}{||C(Dept)||_1}$$

Where

$\widehat{F}(Topic)$ is the normalized funding for each topic within a department,
$P(Topic)$ is the number of Publications for a given topic,
$C(Topic)$ is the number of citations for a given topic,
$C(Dept)$ is the citation for a given department,
$|| \, ||_0$ represents the $l_0 \, norm$, which counts the number of non zero values, and
$|| \, ||_1$ represents the $l_1$ norm, which counts the absolute value of the entries.

The abstract keywords per department were calculated by using NLP algorithms on the dataset containing the abstracts of the publications using unigram, bigram and trigram to find the most frequently occurring sequential tokens. From the sequential tokens, the generic words were removed to formulate a clean corpus which was used to compute the four types of scores defined above.

In order to have a better understanding of these metrics, a graphical representation per department was plotted which has been included in our website and a few of these plots have been shown below.

From Figure 1 (see next page), it can be observed that the abstract word “Structure” from the Department of Chemistry had a publication score of 8, citation score of 310 and an impact score of 1.16 which makes it one of the most influential abstract keyword directly proportional to the department wise publication and citations. The radius of the circle determines the normalized funding generated by the corresponding keyword.

Similarly, from Figure 2, it can be observed that the abstract word “cell” from the Department of Biology, Biomedical Sciences and Genetics had the highest impact score and citation metric.

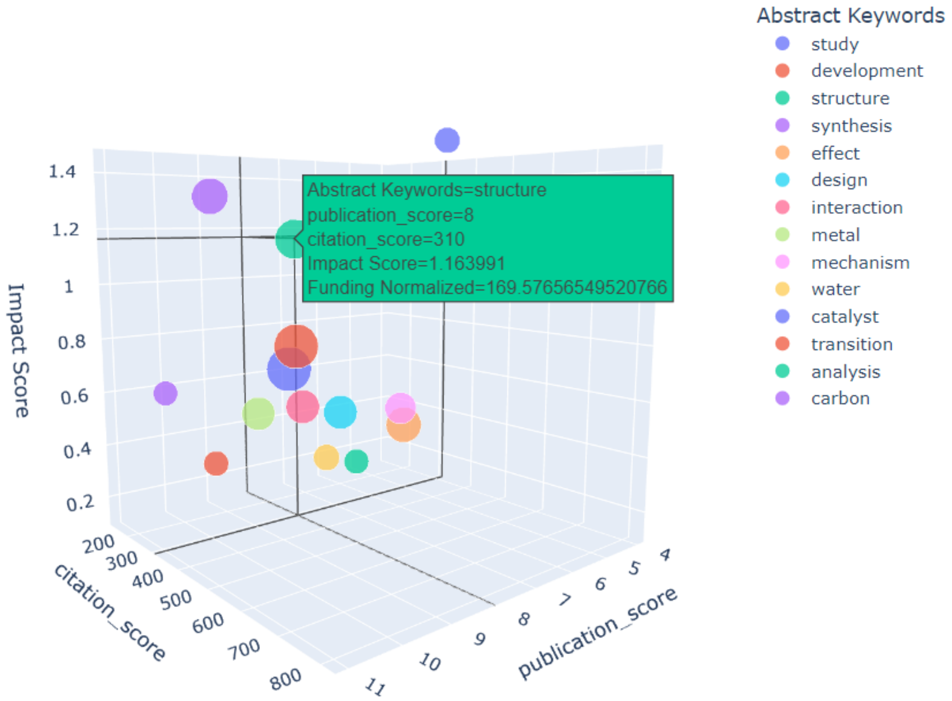


Fig 1. Cumulative Publication score, Impact score and citation score for the Department of Chemistry

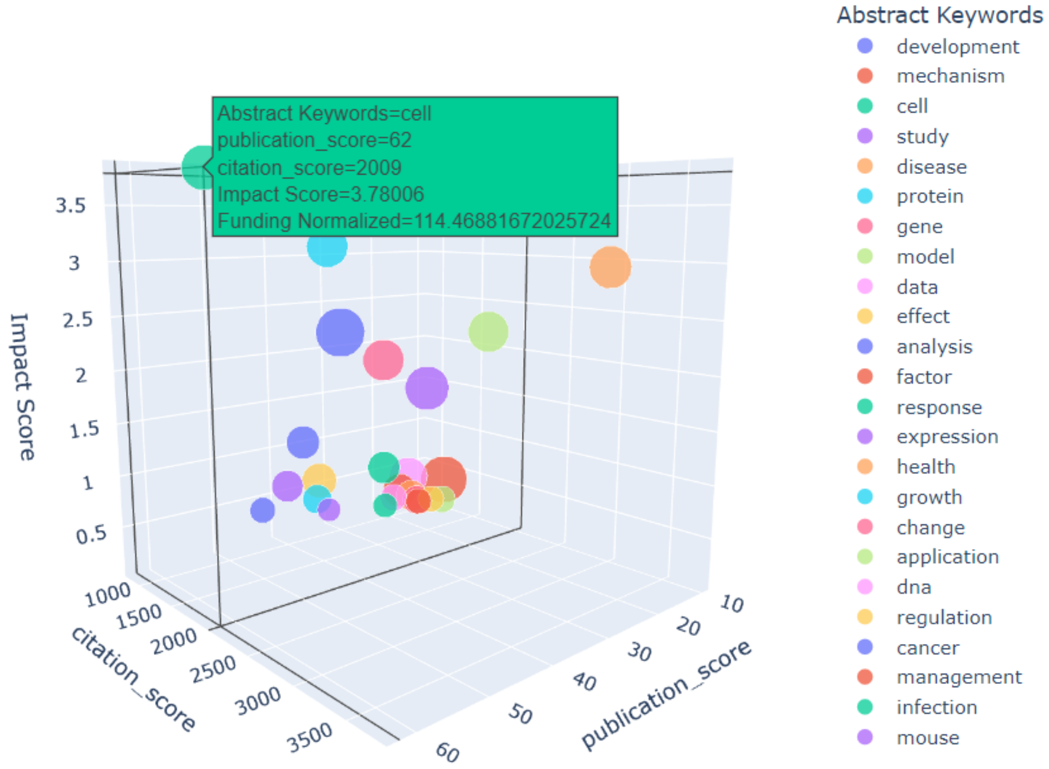


Fig 2. Cumulative Publication score, Impact score and citation score for Biology, Biomedical Sciences and Genetics

# Methodology, Visualization and Interpretation

## Grant Analysis

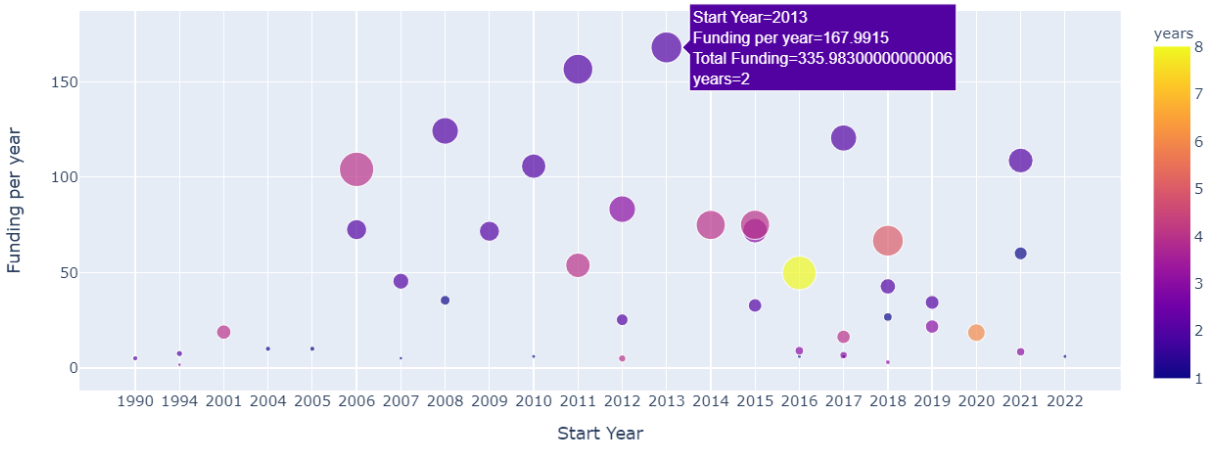
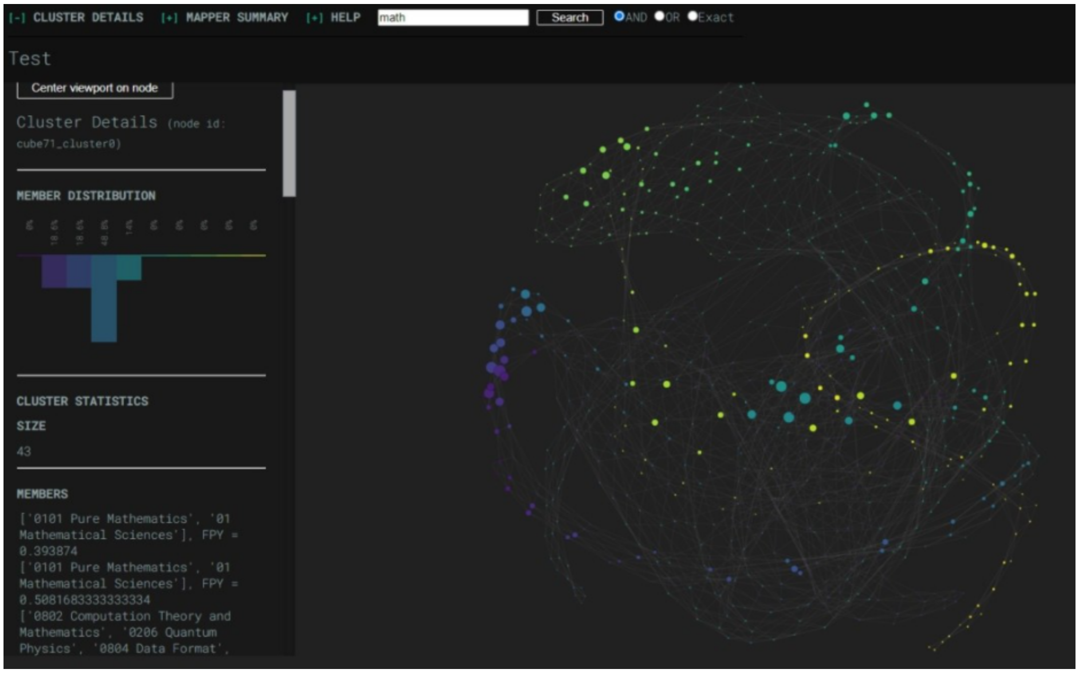


Fig 3. Funding analysis with duration for the Department of History and Archeology

This graph describes the funding received by the Dept of History and Archeology over the years which contains information about the amount of funding received per year and the total amount of funding received for every department. There are various trends which can be observed here, the trivial one being a constant increase in the funding throughout the years. This analysis about the grant acquisition has been done for all the departments and has been presented in an interactive format on the website.



math ● AND ○ OR ○ Exact

Test

Cluster Details (node id: cube71_cluster0)

CLUSTER STATISTICS

SIZE

43

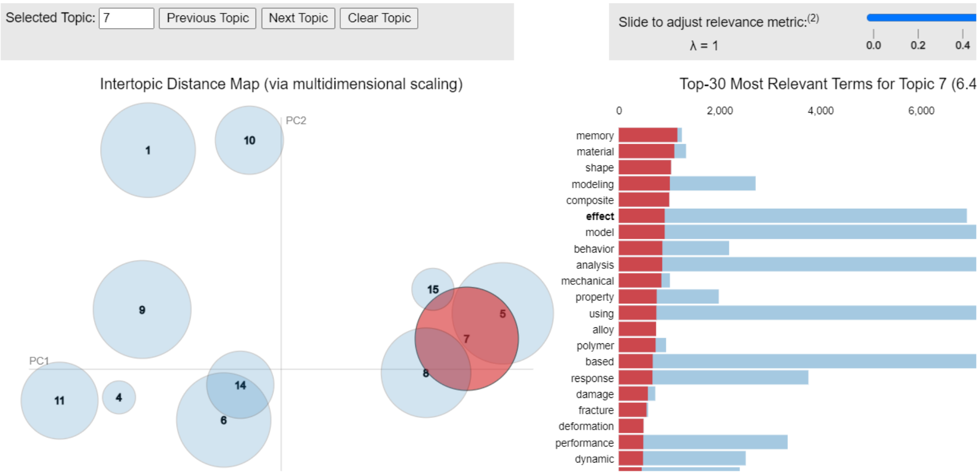
MEMBERS

['0101 Pure Mathematics', '01 Mathematical Sciences'], FPY = 0.393874
['0101 Pure Mathematics', '01 Mathematical Sciences'], FPY = 0.5081683333333334
['0802 Computation Theory and Mathematics', '0206 Quantum Physics', '0804 Data Format',

Fig 4. Department-wise Grant network analysis for Math

Figure 4 describes the clustering for different departments with additional funding per year data. The image above shows the nodes highlighted corresponding to the key “math” in the search box, where the size of the node is proportional to the number of grants available for “math” dept in this case. The mapper algorithm, with Topological Data Analysis, aesthetically plots the data points in 3-dimensions, where we look at the inverse image of projected clusters in the original dataset grouped together.

## Publication Analysis and NLP



Selected Topic: 7

Slide to adjust relevance metric:(2)

λ = 1

Fig 5:Topic modeling for publication abstracts

We have used **Latent Dirichlet allocation** (**LDA**) to extract 15 topics from our corpus, where each topic provides a number of most frequent words within the topic space. In the given image, the topic can be inferred as public health or agriculture research department related publications.

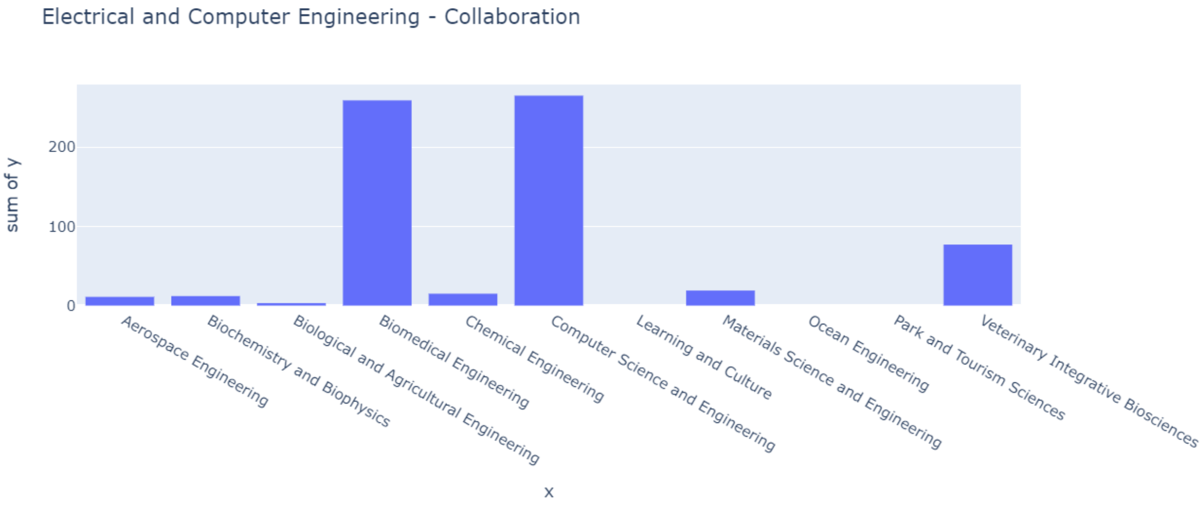
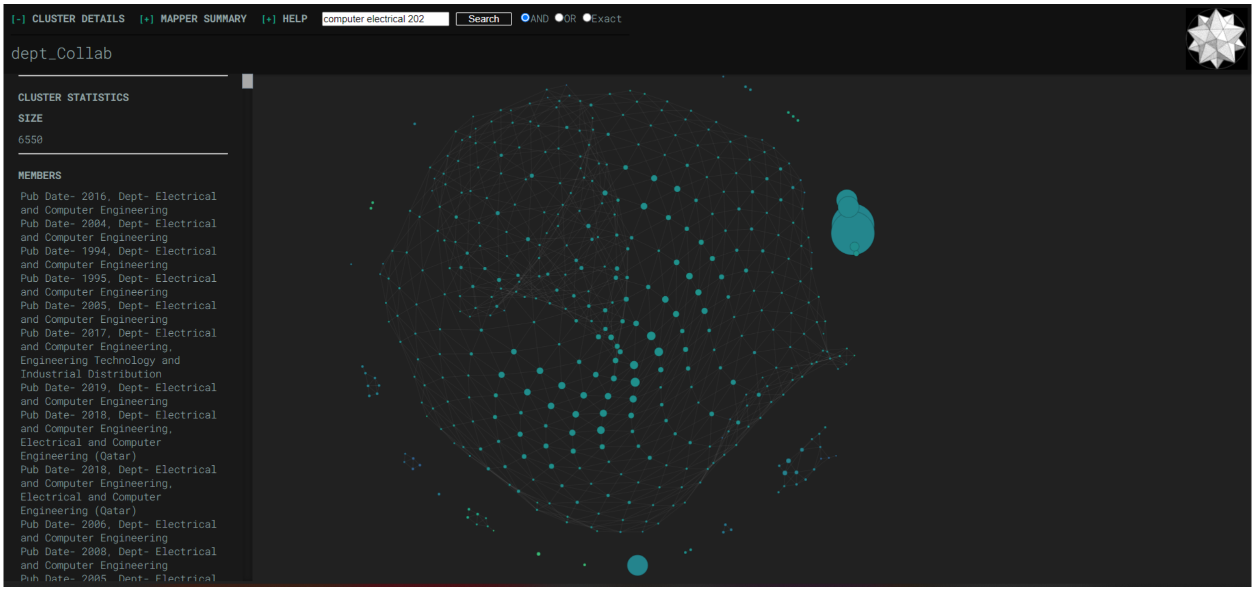


Fig 6: Collaboration between two departments. Both departments are user selected, and the number of the left indicate papers where both the dept collaborate. Selective departments are sampled based on higher publication numbers as well as at least one collaboration with some other dept.

Figure 6 represents the existing collaboration between departments along with the insight related to the similarity between them based on the publication data obtained from altmetric website. We can select a department and look at the common publications they have with some other departments. The departments are sampled based on high publication counts and non zero collaboration.



computer electrical 202

● AND ○ OR ○ Exact

dept_Collab

CLUSTER STATISTICS

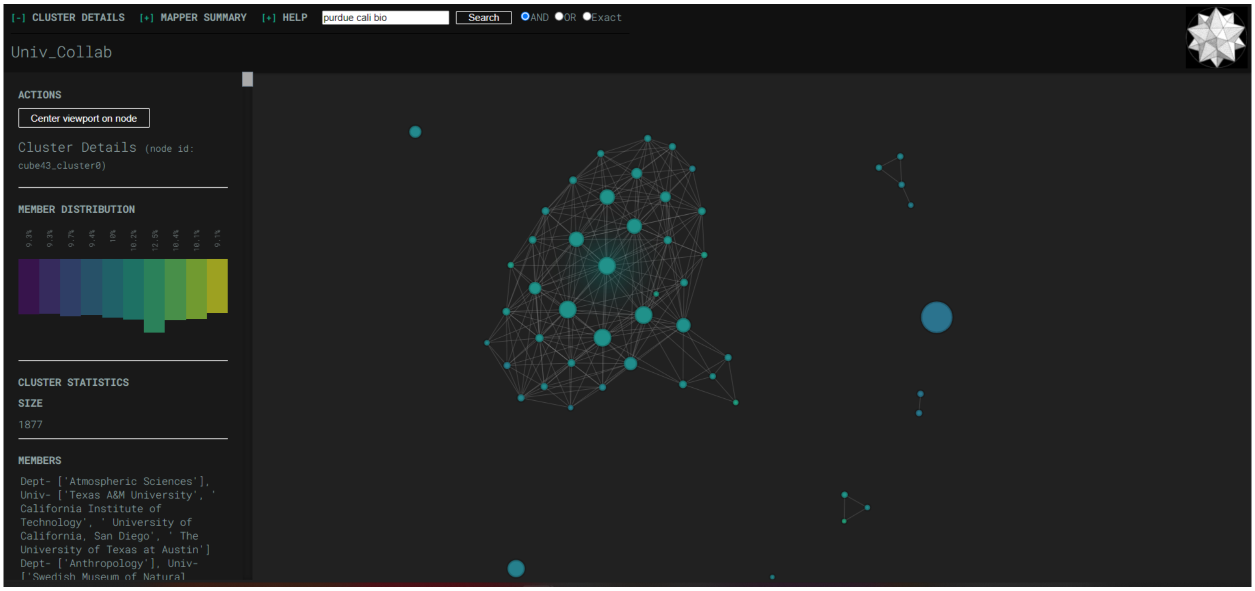
SIZE

6550

MEMBERS

```
Pub Date- 2016, Dept- Electrical
and Computer Engineering
Pub Date- 2004, Dept- Electrical
and Computer Engineering
Pub Date- 1994, Dept- Electrical
and Computer Engineering
Pub Date- 1995, Dept- Electrical
and Computer Engineering
Pub Date- 2005, Dept- Electrical
and Computer Engineering
Pub Date- 2017, Dept- Electrical
and Computer Engineering,
Engineering Technology and
Industrial Distribution
Pub Date- 2019, Dept- Electrical
and Computer Engineering
Pub Date- 2018, Dept- Electrical
and Computer Engineering,
Electrical and Computer
Engineering (Qatar)
Pub Date- 2018, Dept- Electrical
and Computer Engineering,
Electrical and Computer
Engineering (Qatar)
Pub Date- 2006, Dept- Electrical
and Computer Engineering
Pub Date- 2008, Dept- Electrical
and Computer Engineering
Pub Date- 2005, Dept- Electrical
```

Fig 7: Collaboration between Electrical and Computer science dept, in the last two years. Data input from the user can be entered in the search box. This network graph obtained through topological data analysis provides a faster method to look at the number of publications for different departments along with a filter to look at the range of years one wants to focus on. With some tweaks, one can see the trend in publication data for different departments throughout the years.



purdue cali bio

● AND ○ OR ○ Exact

Univ_Collab

ACTIONS

Cluster Details (node id: cube43_cluster0)

MEMBER DISTRIBUTION

CLUSTER STATISTICS

SIZE

1877

MEMBERS

```
Dept- ['Atmospheric Sciences'],
Univ- ['Texas A&M University', '
California Institute of
Technology', ' University of
California, San Diego', ' The
University of Texas at Austin']
Dept- ['Anthropology'], Univ-
['Swedish Museum of Natural
```

Fig 8: Collaboration between different universities. Interactive with an option to look for different universities and also the department. Example of collaboration between Purdue, UC in bio related topics.

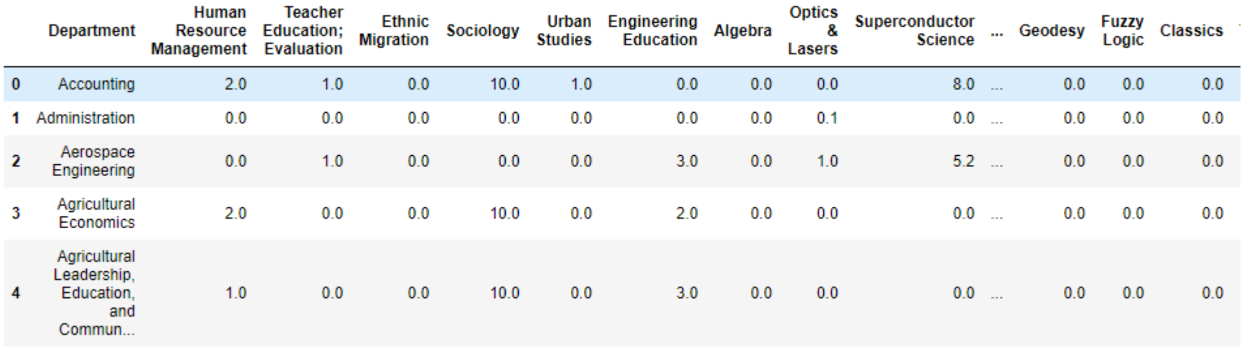
Another interesting way to look at an interactive graph, where one can look up universities and see whether there are collaborations already. The university data has been projected, clustered and the mapped back to the original data to find a more optimal group structure with additional information about the connectivity of different nodes/data points.

## Potential of Inter-Department Collaboration

The Texas A&M University System consists of many departments working at multiple places around the globe. Due to the high number of research and study at these departments, it is possible that similar studies are carried out without much interaction between the departments or the researchers. Such collaborations if possible will lead to increase in the research networks, reduced expenditure, and access to increased publications thus resulting in better research productivity.

It is thus important to quantify the potential of research collaborations among the researchers. We have come up with an algorithm that finds out the similarity metric for different department combinations. The algorithm works on keywords extracted from the publication data; 554 such keywords or scientific subdisciplines have been defined across publication for every researcher. The extracted keyword/subdiscipline data is readily available for 4840 individuals in the CSV file format and was web-scrapped from the Scholar@Tamu Map of Science website. Each of these 554 subdisciplines are assigned numerical values equal to the number of publications made by the researcher.

The number of combinations of individual researchers from different departments was way too high to compute the potential collaboration score of each combination. We thus group the entire data into departments resulting in fewer combinations to check.



| | Department | Human Resource Management | Teacher Education; Evaluation | Ethnic Migration | Sociology | Urban Studies | Engineering Education | Algebra | Optics & Lasers | Superconductor Science | ... | Geodesy | Fuzzy Logic | Classics |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Accounting | 2.0 | 1.0 | 0.0 | 10.0 | 1.0 | 0.0 | 0.0 | 0.0 | 8.0 | ... | 0.0 | 0.0 | 0.0 |
| 1 | Administration | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | ... | 0.0 | 0.0 | 0.0 |
| 2 | Aerospace Engineering | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 3.0 | 0.0 | 1.0 | 5.2 | ... | 0.0 | 0.0 | 0.0 |
| 3 | Agricultural Economics | 2.0 | 0.0 | 0.0 | 10.0 | 0.0 | 2.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 |
| 4 | Agricultural Leadership, Education, and Commun... | 1.0 | 0.0 | 0.0 | 10.0 | 0.0 | 3.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 |

Table 1. Number of publications per associated discipline in the Department eligible for potential collaboration

As seen from Table 1, each department is given a value for the sub-disciplines and we have analyzed 408 such departments. Each row corresponding to the department can be considered as a vector of length 554 with sub-disciplines as the elements in the vector characterizing the department.
Next, a similarity score is generated for every pairwise combination of the department and in this case, we use the cosine similarity. Higher the cosine similarity score between two departments, higher is their potential to work together on various subdisciplines.

The cosine similarity is defined as follows:

$$\text{cosine similarity} = S_C(A,B) := \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}},$$

where A and B are the rows of table or vectors of subdiscipline keywords of length 554.

## Findings from this Analysis

A total of 83028 ($^{408}C_2$) comparisons were made and we could see that various departments were exclusively working on similar subdisciplines. We could thus make recommendations to these departments to collaborate if they don't already. The following plots show the departments that have high potential of collaboration, the plots also show the subdisciplines in which the departments should be collaborating. We visually analyze the graphs to analyze which is a better approach for the subdiscipline representation.

| Department 1 | Department 2 | Potential Collaboration Score |
|---|---|---|
| Center for Microencapsul and Drug Delivery | Irma Lerma Rangel College of Pharmacy | 0.9953 |

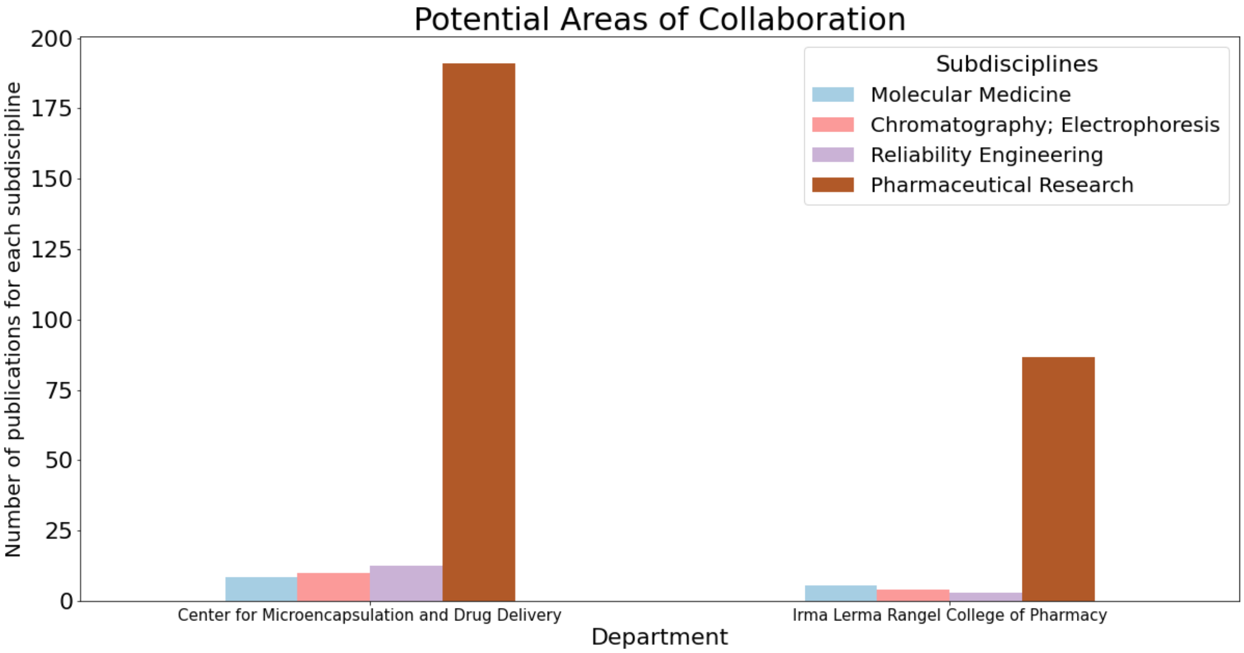


Fig 7. Collaboration potential between Center for Microencapsulation and College of Pharmacy

As seen in Fig 7, the cosine similarity denoting the potential collaboration score between Center for Microencapsul and Drug Delivery, and Irma Lerma Rangel College of Pharmacy is high. These departments can collaborate on the subdisciplines of Molecular Medicine, Chromatography, Reliability Engineering and Pharmaceutical Research. We also see that both the departments specialize in Pharmaceutical Research due to high publication research output.

Another example of possible collaboration is as follows.

| **Department 1** | **Department 2** | **Potential Collaboration Score** |
|---|---|---|
| Political Science | Public Service and Administration | 0.8590 |

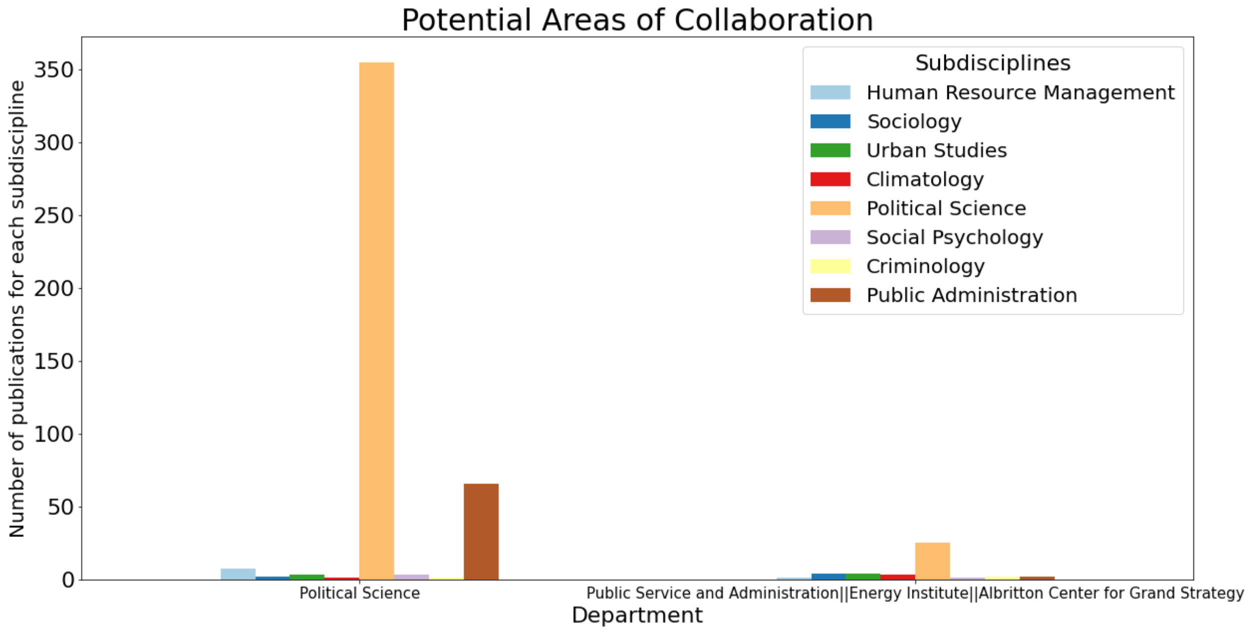


Fig 8. Collaboration potential between Department of Political Science and Department of Public Administration

Thus, as per the methodology we have followed we recommend that the departments Political Science, and Public Service and Administration work together on the subdisciplines of Human Resource Management, Sociology, Urban Studies, etc.

# Conclusions and Recommendation

By looking at the collaborative potential score, there is a high potential for multiple departments to work on certain disciplines together, which can possibly lead to an increase in the overall research productivity, reduced research expenditure, and higher overall research network.

Impact score created through utilization of citation scores, publication data and grant information, could help us sample better topics to enhance the overall research conducted in various departments. A substantially sound conjecture can be”

*“Increasing the impact score overall for different topics in a department could direct the department towards an overall growth and improve the rank of the department (as well as the university) across the globe”.*

A few things we can recommend are: A combination of an altmetric score and impact score, depending on the public disclosure of different departments can lead to a more efficient research environment inside the department.
Through potential collaboration scores, the research output for several topics can be increased with a decrease in the time taken for publications, and added perspective to the research for different departments.