A Needle in a Data Haystack (67978) - Final Project
Analyzing and Predicting Trends in Academic Research

Team Members:
Noa Ben Gallim, noa.bengallim@mail.huji.ac.il, noabengallim
Itamar Edelstein ,itamar.edelstein@mail.huji.ac.il, ita.edel
Idan Hippach, idan.hippach@mail.huji.ac.il, idanhippach

Problem Description

The primary goal of our project is to analyze and predict trends in academic research over time, focusing on identifying trending topics and mapping collaboration patterns among researchers across different fields. This project leverages data from the Microsoft Academic Graph (MAG) to identify the most popular research topics, track their evolution, and visualize these trends dynamically. Additionally, we aim to investigate co-authorship patterns to uncover interdisciplinary collaboration networks, revealing how different research fields are interconnected through joint efforts.

Data

We utilized the MAG paper_schema, which is a comprehensive database of academic papers, authors, organizations, and technical metadata. This includes critical data about each paper's authorship, keywords, organization affiliation, and the venue of publication.

- **Data Forma**t: Multiple .txt files, each approximately 10GB in size, with rows in JSON format. The total dataset contains around 12 million rows.
- Source: The data was downloaded from the AMiner website.

Our Solution

Preprocessing

To make the dataset manageable, we preprocessed each .txt file by performing the following steps:

1. **Column Removal**: Removed columns containing unnecessary technical metadata, such as the venue of presentation and page count.
2. **Null Handling**: Removed rows with missing values or duplicated entries to ensure data quality.
3. **Time Filtering**: Filtered the dataset to only include papers published only between 2000-2019.
4. **Author Filtering**: Removed entries lacking author or affiliation information, ensuring that only papers with complete author metadata (including university affiliation and university code) were retained.

The cleaned and filtered data from each .txt file was then merged into a single .csv file containing 12 million rows.

Key Features Retained

The final dataset includes the following key features:

- **id**: Unique identifier for each paper.

- **title**: Title of the paper.
- **authors**: Metadata about each paper's authors and their affiliated universities.
- **year**: Year of publication.
- **n_citations**: Number of times the paper has been cited.
- **fos** (Fields of Study): Research areas covered by the paper, indicating the subject areas.

Data Analysis

We utilized several key techniques to analyze the fields of study (fos) and collaborations across academic papers:

1. **TF-IDF Representation**: We applied TF-IDF to represent each paper's fields of study. This allowed us to capture the importance of terms within each paper's topic, which was then used for clustering.
2. **K-means Clustering**: Using the TF-IDF representations, we applied K-means clustering to group papers into broader subjects. After experimenting with various values of K, we selected the optimal number of clusters by evaluating the results and visualizing the terms with word clouds.
3. **Community Detection**: To explore interdisciplinary collaborations, we used Louvain community detection on the co-authorship networks. This helped identify key research communities and the strength of collaborations between fields, which were then plotted over time to show their evolution. The labels for the communities were taken from the fields that appeared the most in each community.

Evaluation

Evaluation Criteria

We evaluated the success of our approach using three key criteria:

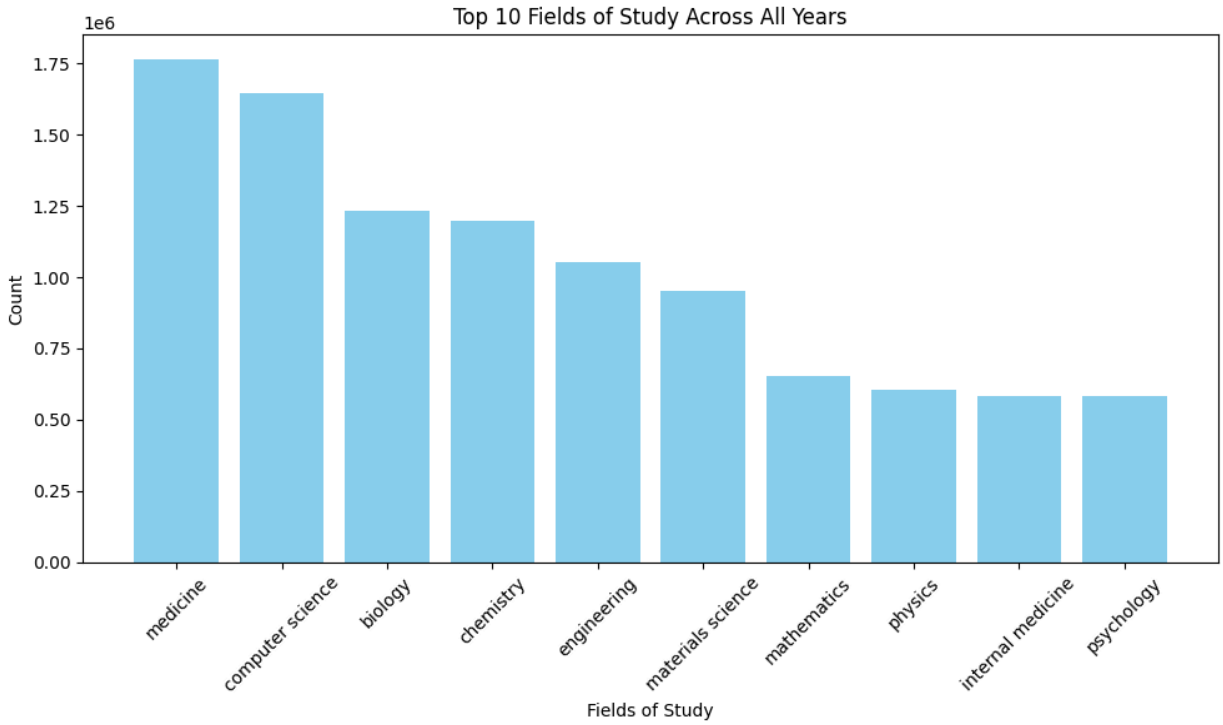
1. **Clustering Accuracy**: We assessed the quality of the K-means clustering by examining the coherence of topics in each cluster, using word clouds to visualize and interpret the general subjects.
2. **Publication Trends**: The success of our visualizations (bar and line plots) was measured by how well they aligned with established trends in research. We checked whether the identified top fields reflected known global trends.
3. **Interdisciplinary Collaborations**: The effectiveness of our community detection was evaluated by identifying significant collaborations between fields and tracking these collaborations over time.

Setup

Our analysis involved processing vast datasets across multiple machines whilst optimizing performance using distributed processing techniques to get our results. We made sure that the project was completed efficiently despite the large dataset size.

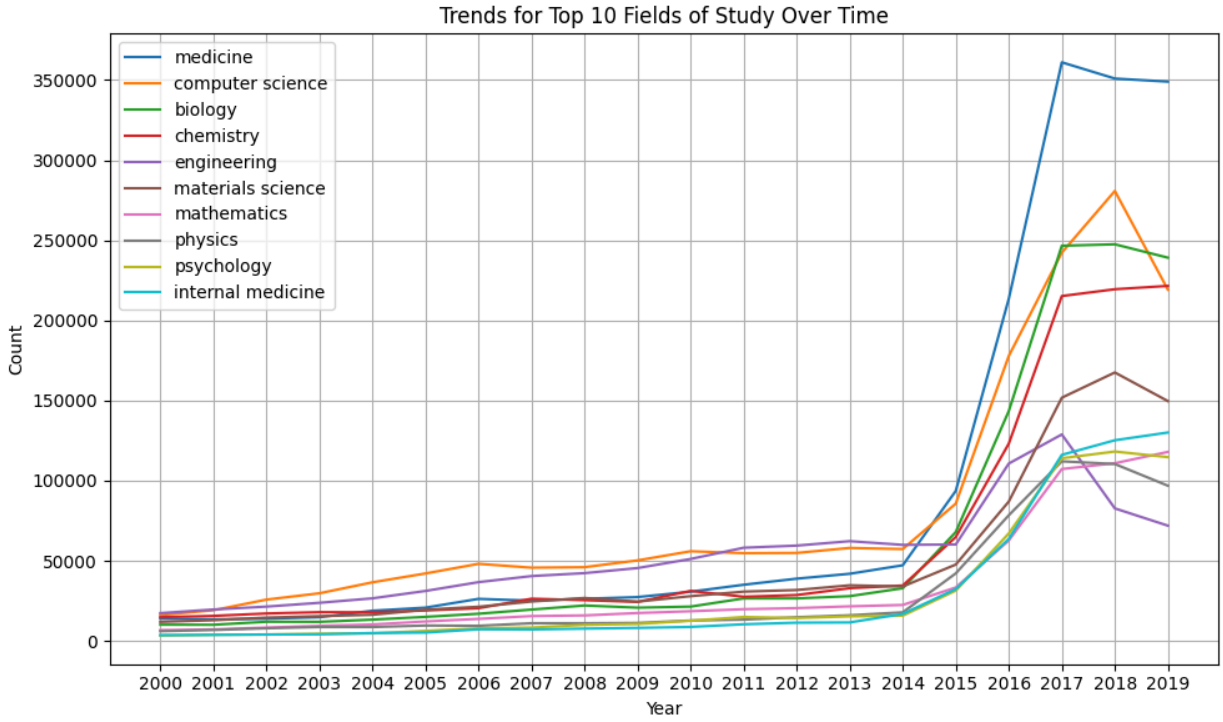
Visualization

1. Top Fields of Study:



The bar graph shows the top ten fields of study across all years, highlighting Medicine as the most dominant field with nearly 1.75 million publications. Following Medicine are Computer Science and Biology, which also show significant research output. Fields like Chemistry, Engineering, and Materials Science are also highly represented. This bar chart effectively conveys the distribution of publications across different research areas, making it easy to compare the dominance of certain fields. Medicine's prominence suggests that healthcare-related research has been a consistent priority over time, while the strong presence of technology-driven fields like Computer Science reflects the growing importance of technological advancements in academic research.

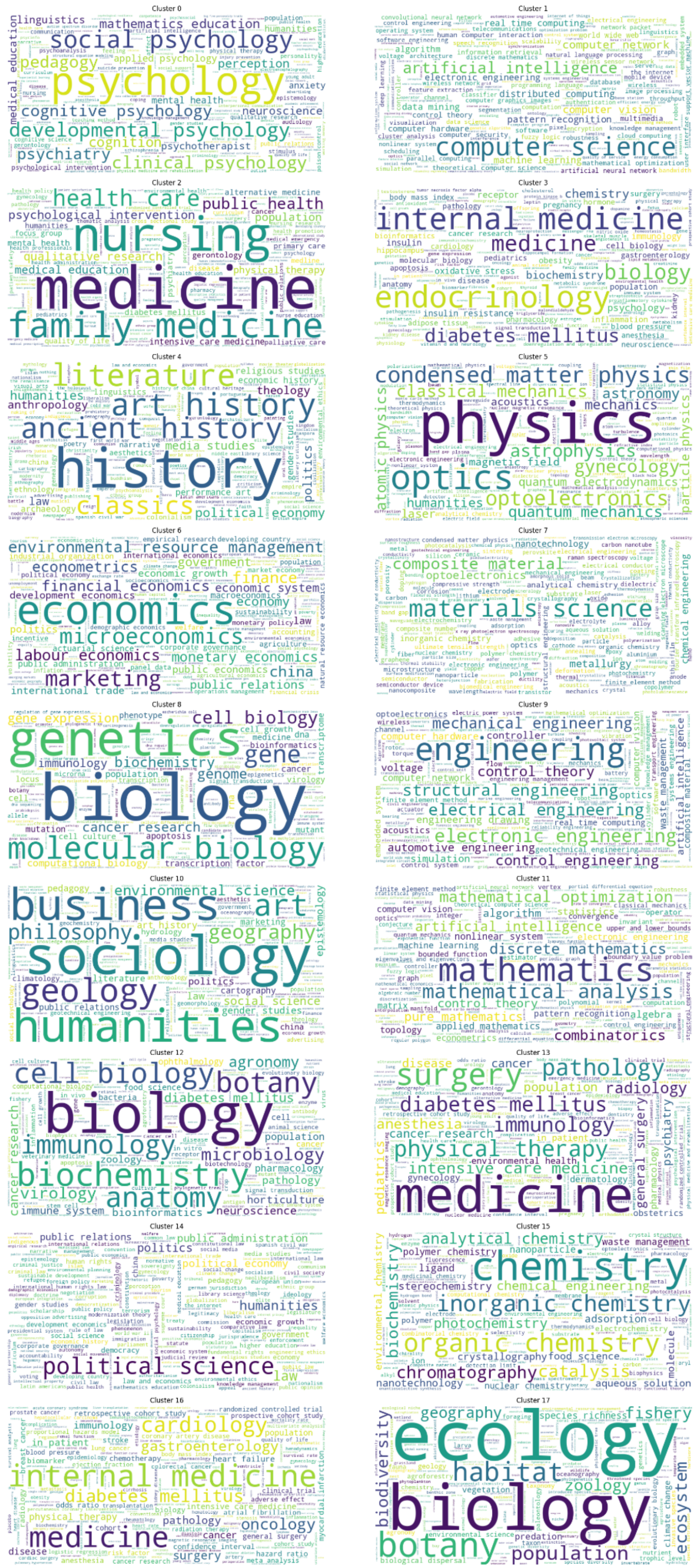
2. Trends Over Time:



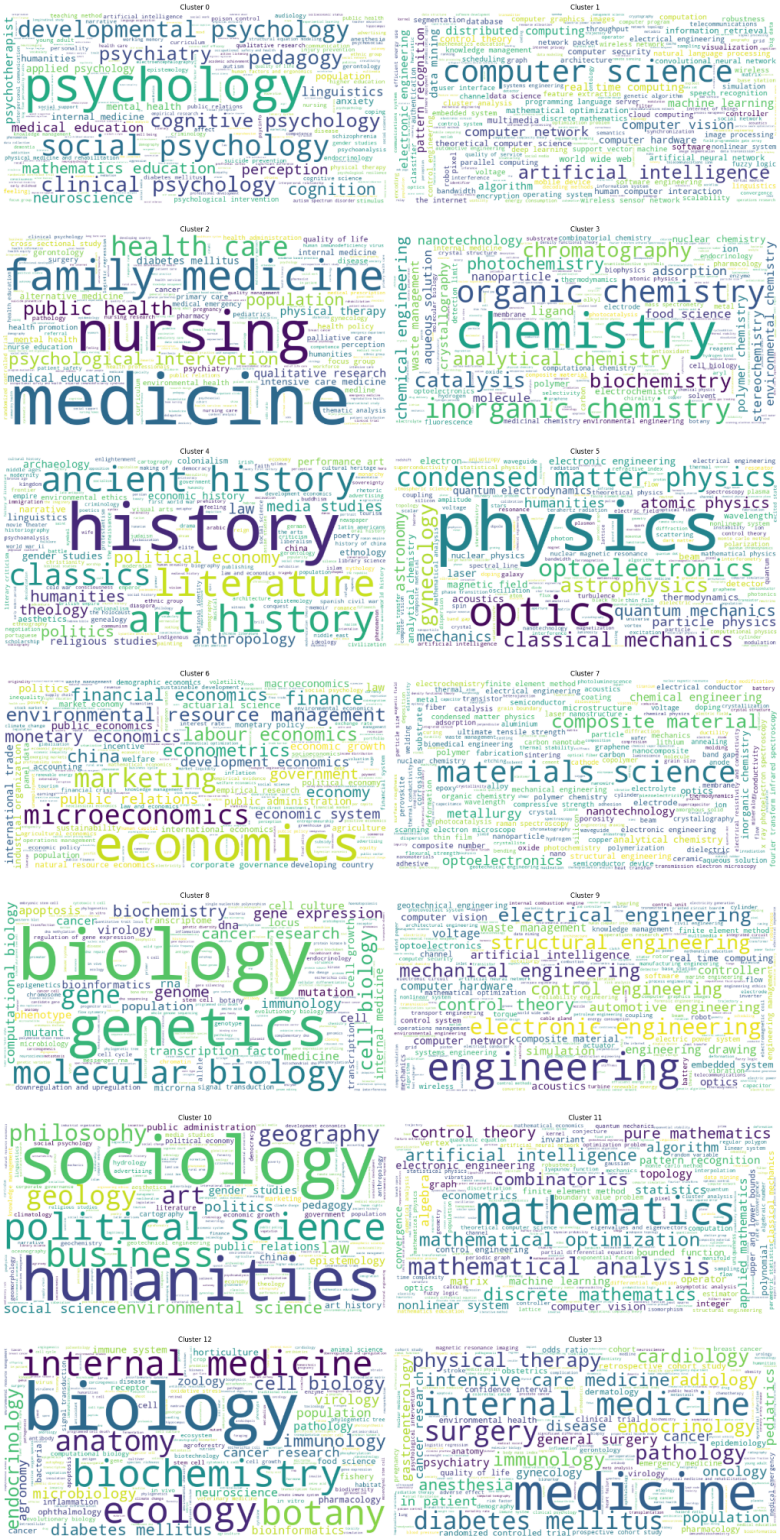
The line plot visualizes the trends in research output across the top ten fields from 2000 to 2019. The plot shows a clear upward trajectory for all fields, with notable spikes around 2015 to 2018. Medicine again stands out as the field with the highest number of publications, but Computer Science exhibits a particularly sharp rise in output during this period, indicating a rapid increase in technological research. Fields such as Biology and Chemistry have also gained significant attention over the years, showing strong growth in research activity. Additionally, Materials Science demonstrates a steady rise.

3. K-Means Clustering:

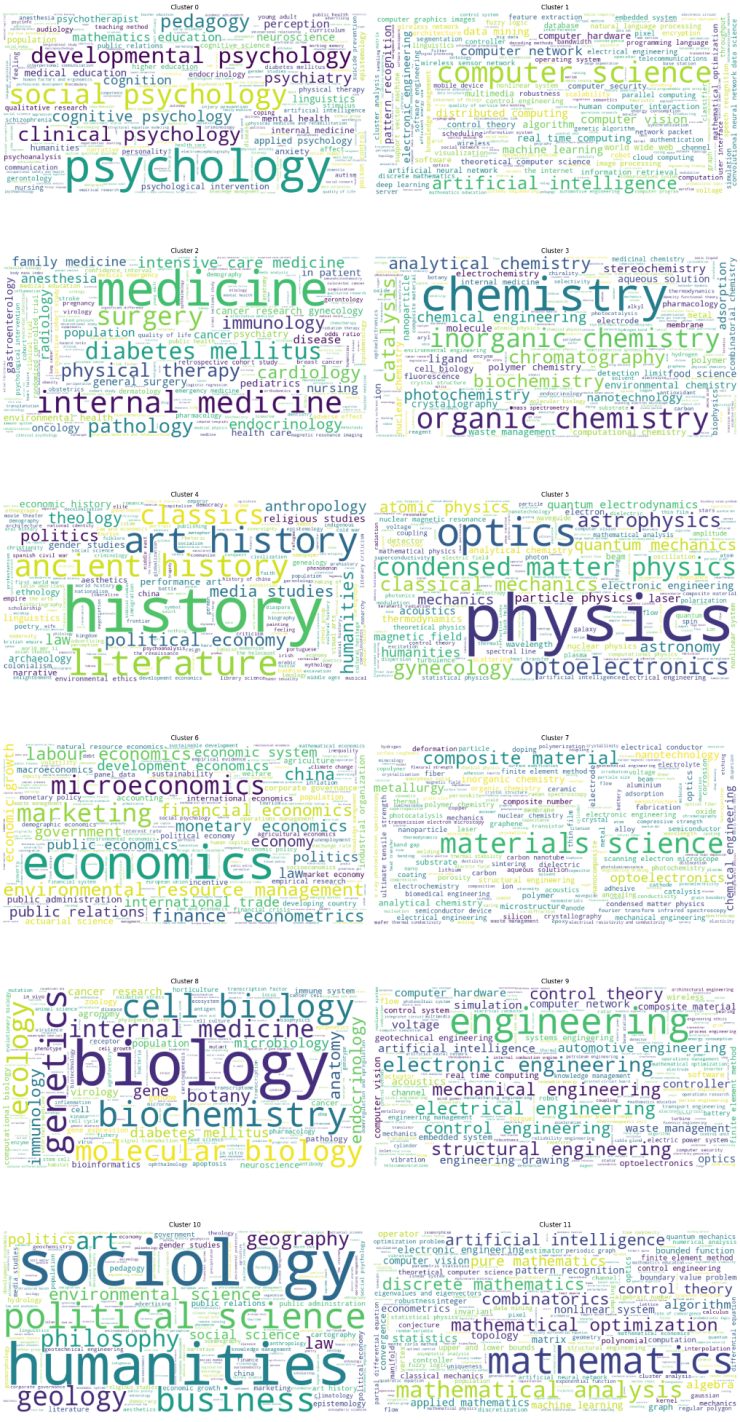
In the K-means clustering results with **K=18**, several clusters overlap or represent closely related fields. For instance, clusters 2, 3, 13, and 16 all focus on medicine and its subfields, while clusters 8, 12, and 17 revolve around various aspects of biology.


Cluster 0
psychology
social psychology
developmental psychology
cognitive psychology
clinical psychology
psychiatry
pedagogy
mathematics education
neuroscience
Cluster 1
computer science
artificial intelligence
computer network
real time computing
distributed computing
Cluster 2
medicine
nursing
family medicine
health care
public health
Cluster 3
internal medicine
endocrinology
diabetes mellitus
medicine
biology
Cluster 4
history
art history
ancient history
literature
classics
Cluster 5
physics
optics
condensed matter physics
astrophysics
optoelectronics
quantum mechanics
Cluster 6
economics
microeconomics
environmental resource management
financial economics
econometrics
marketing
labour economics
monetary economics
Cluster 7
materials science
composite material
metallurgy
chemical engineering
Cluster 8
biology
genetics
molecular biology
cell biology
gene
Cluster 9
engineering
electronic engineering
electrical engineering
structural engineering
mechanical engineering
control engineering
control theory
Cluster 10
sociology
humanities
business
art
geology
philosophy
geography
Cluster 11
mathematics
mathematical analysis
mathematical optimization
discrete mathematics
artificial intelligence
combinatorics
Cluster 12
biology
biochemistry
cell biology
immunology
botany
anatomy
microbiology
Cluster 13
medicine
surgery
pathology
diabetes mellitus
physical therapy
immunology
Cluster 14
political science
politics
public relations
public administration
humanities
law
Cluster 15
chemistry
organic chemistry
inorganic chemistry
analytical chemistry
chromatography
catalysis
Cluster 16
internal medicine
medicine
cardiology
diabetes mellitus
gastroenterology
oncology
Cluster 17
ecology
biology
botany
habitat
population
biodiversity
ecosystem

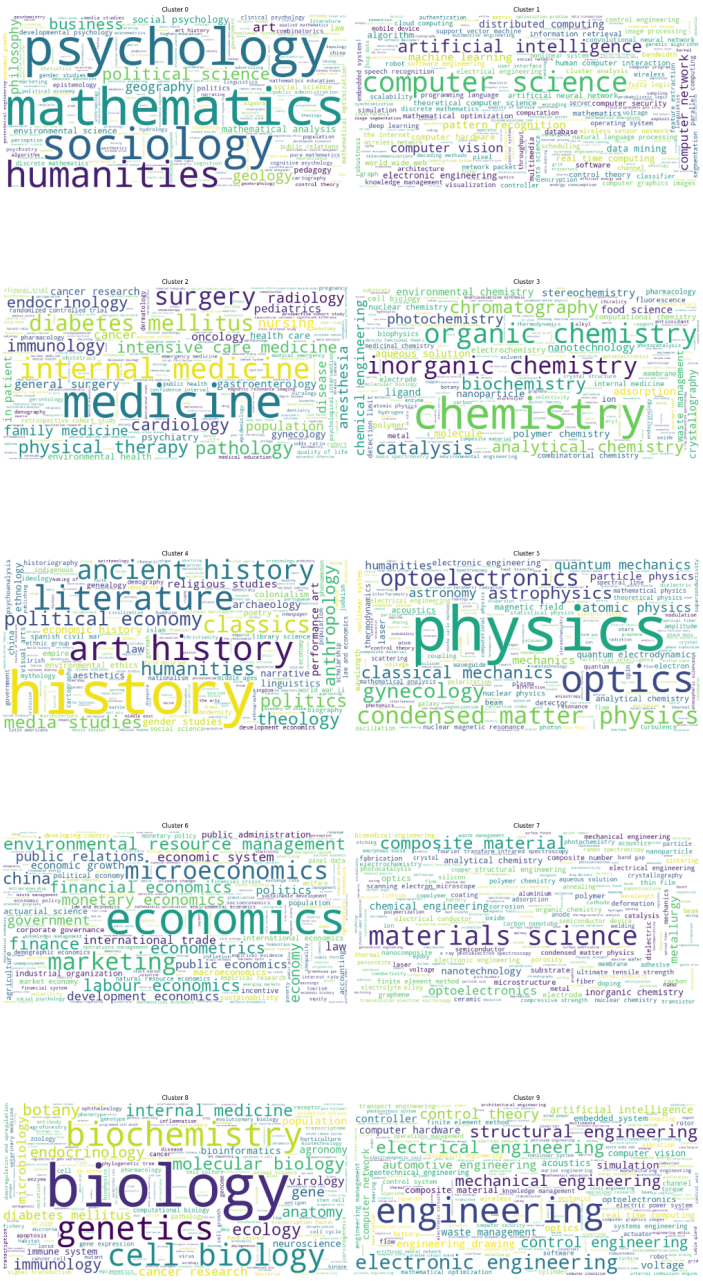
Reducing the number of clusters to **K=14** did not fully resolve this issue, as some overlaps persist in the same areas.



With **K=12**, the separation improves, particularly with biology and medicine now forming distinct clusters. Other fields, such as psychology, computer science, and physics, also remain well-separated.

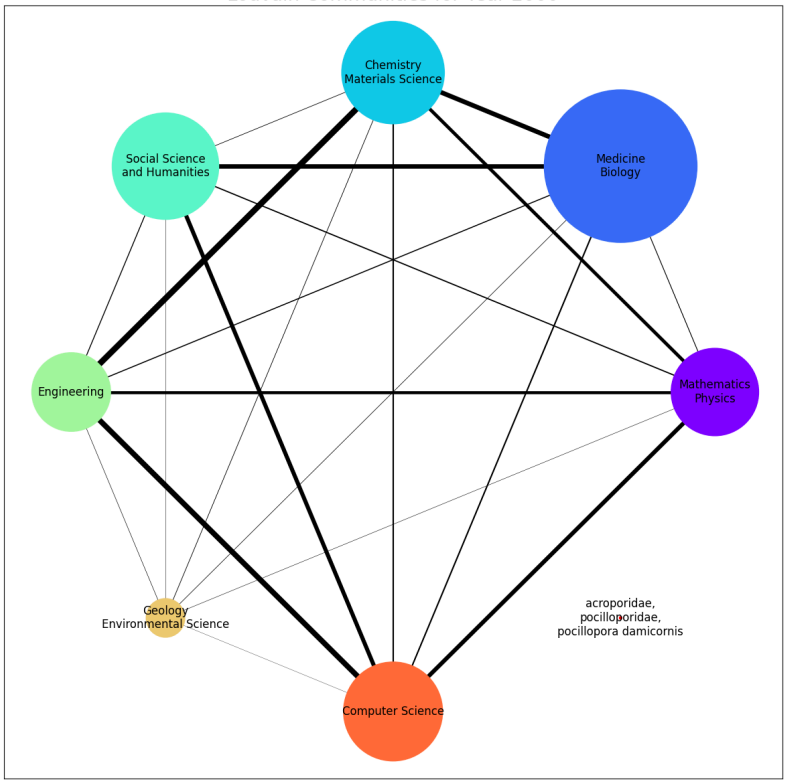


However, when clustering is reduced to **K=10**, some unrelated fields begin to merge, such as psychology and mathematics in cluster 0, indicating that this level of clustering is less effective as it mixes disciplines that should remain distinct.

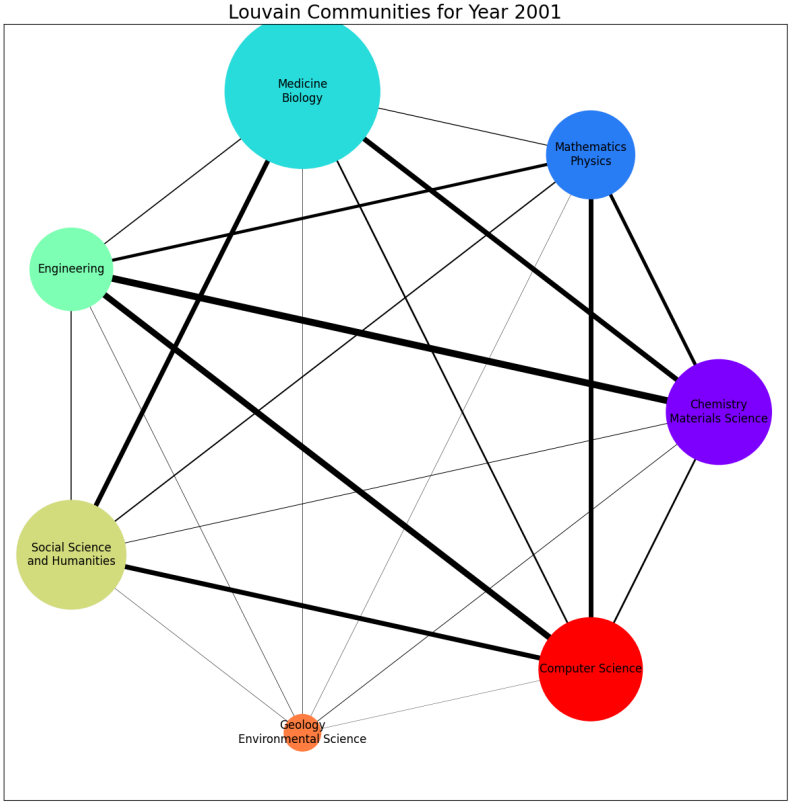


4. Community Detection:

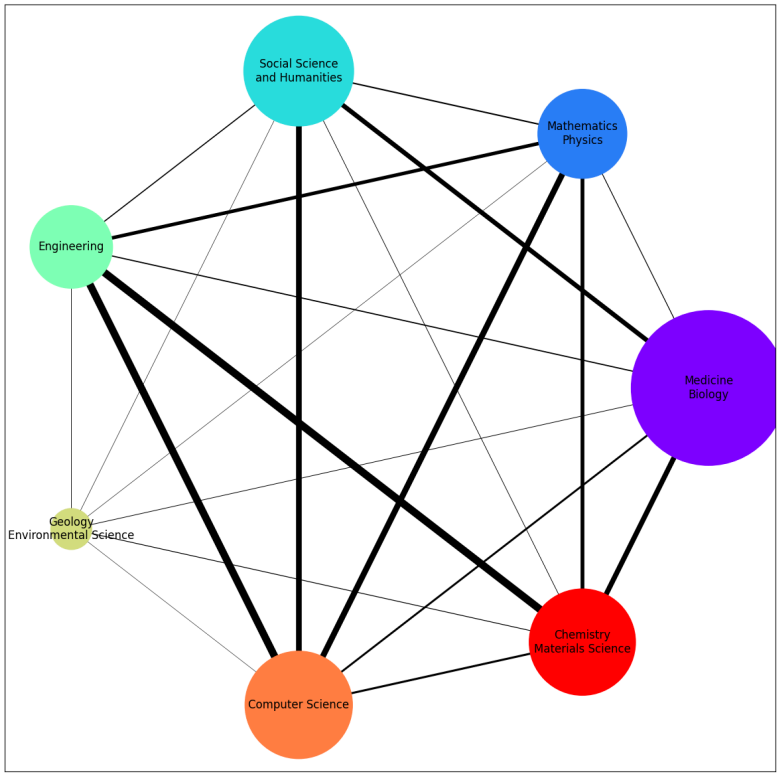
Louvain Communities for Year 2000



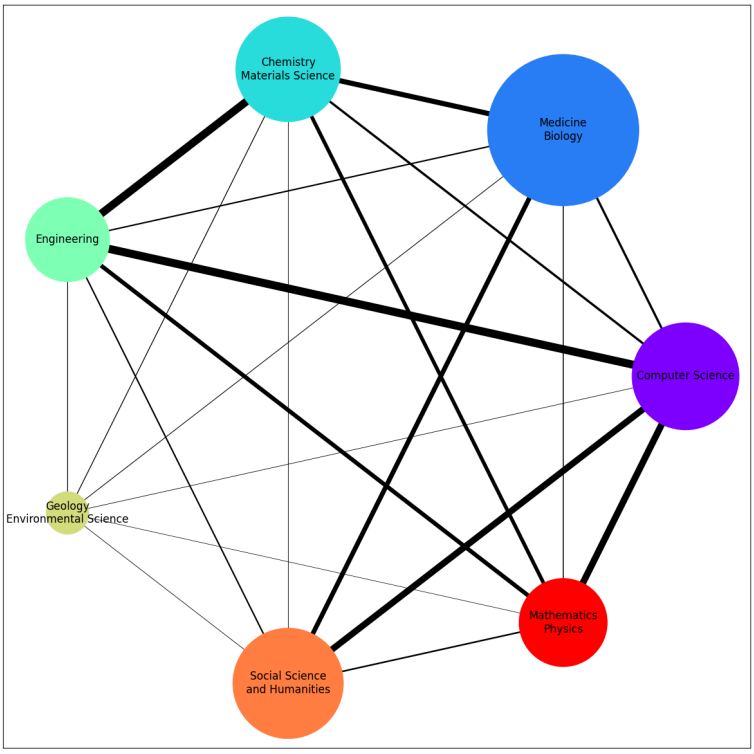
Louvain Communities for Year 2001



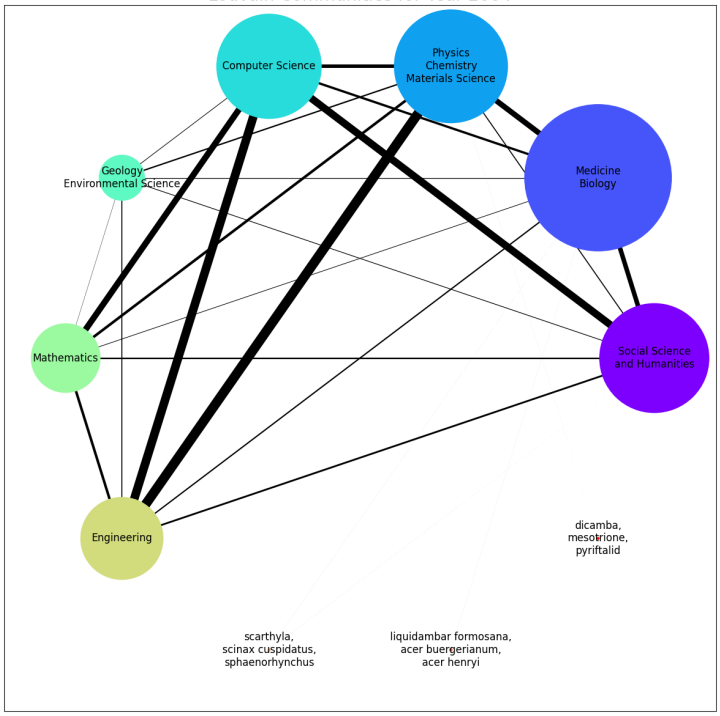
Louvain Communities for Year 2002



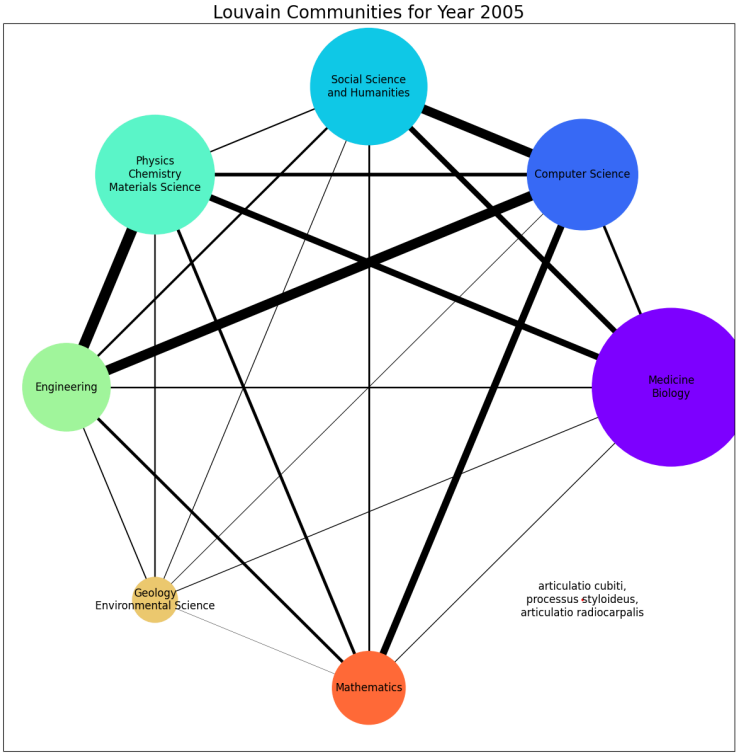
Louvain Communities for Year 2003



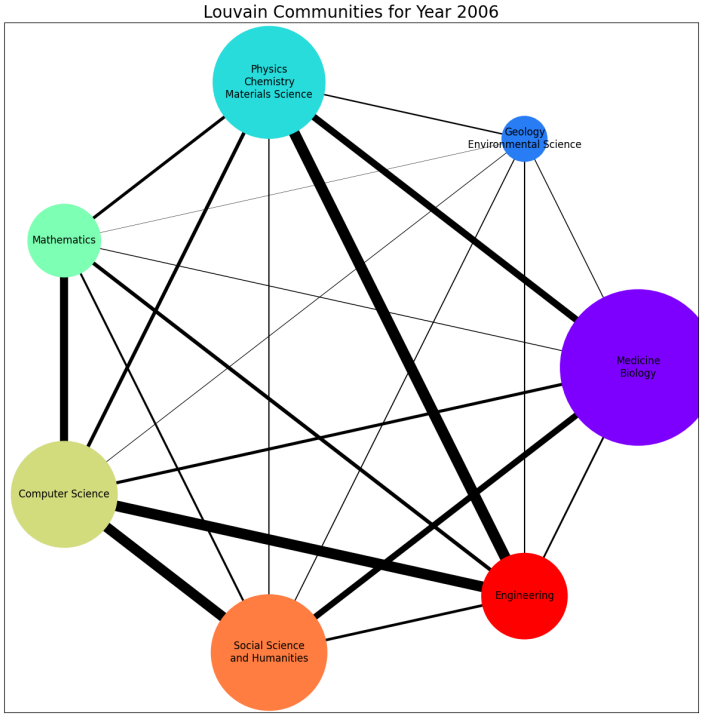
Louvain Communities for Year 2004



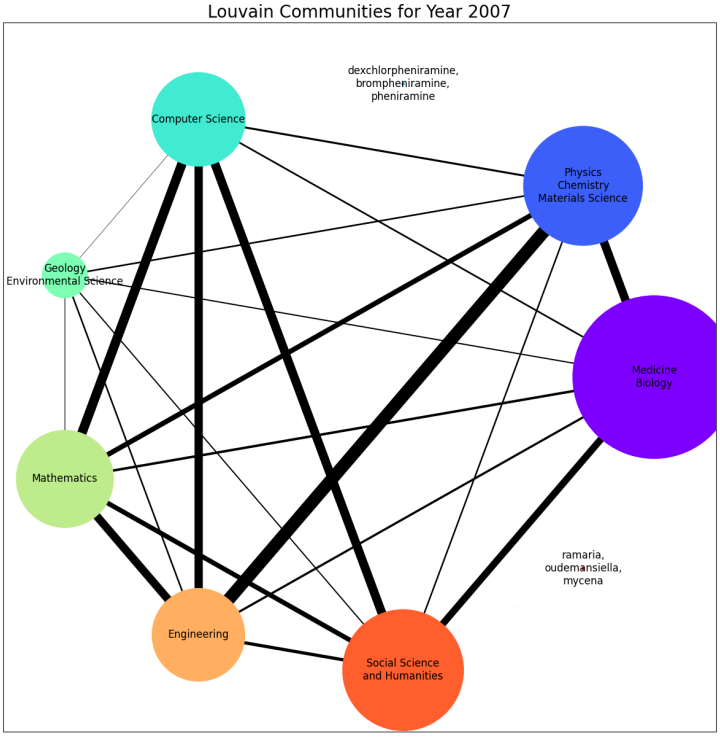
Louvain Communities for Year 2005



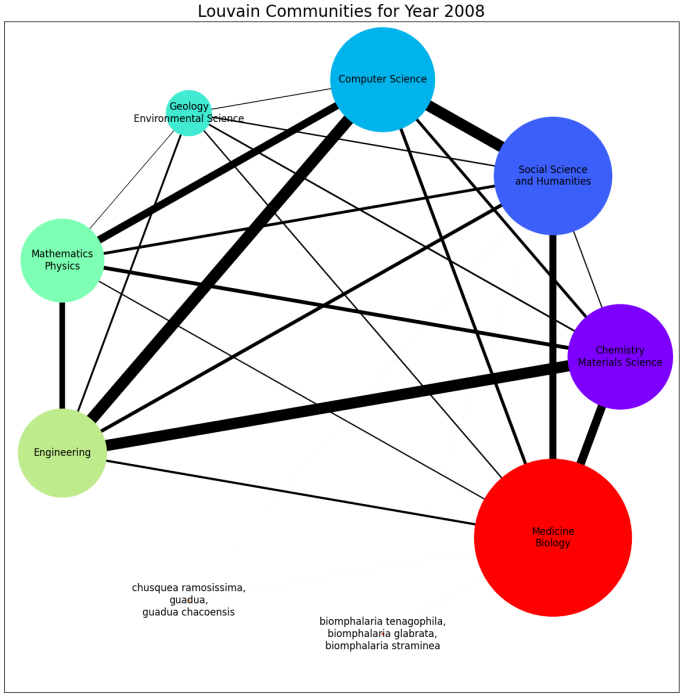
Louvain Communities for Year 2006



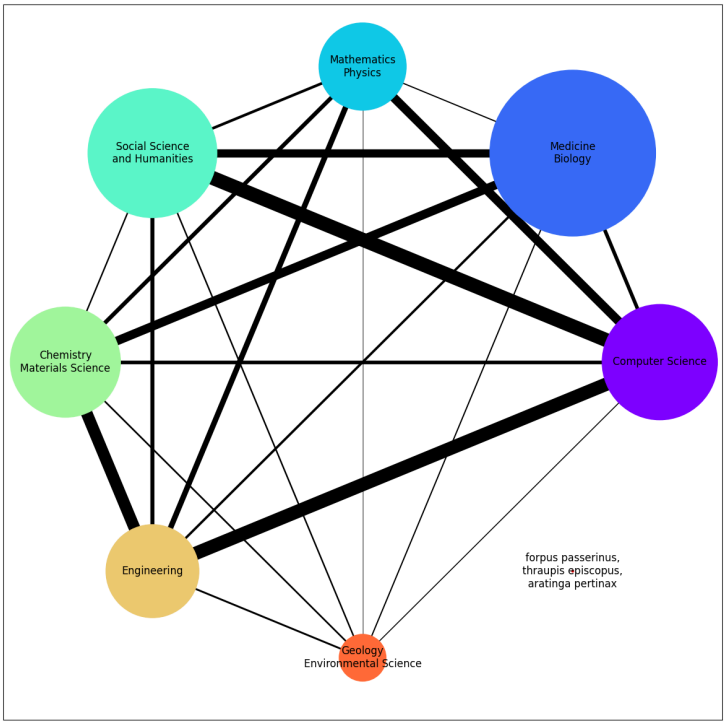
Louvain Communities for Year 2007



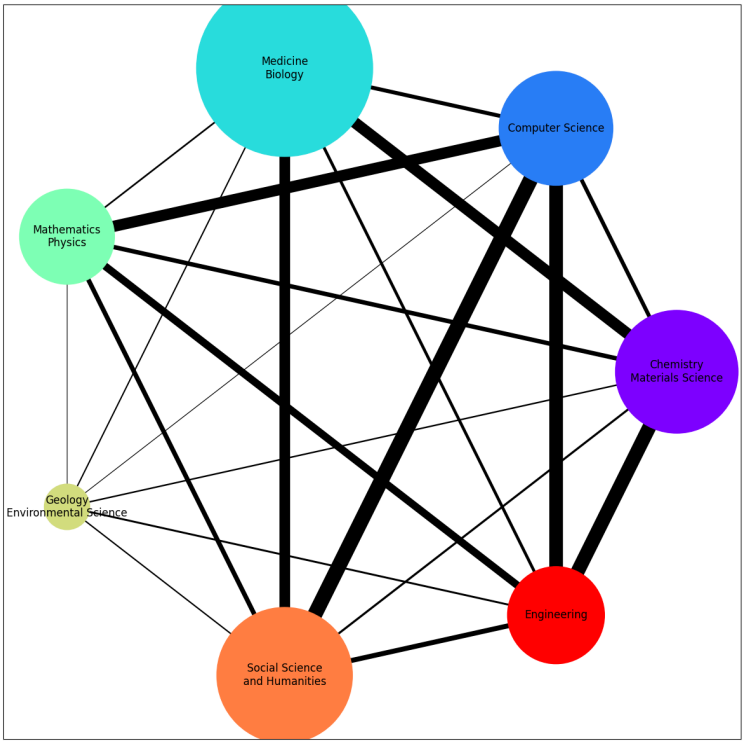
Louvain Communities for Year 2008



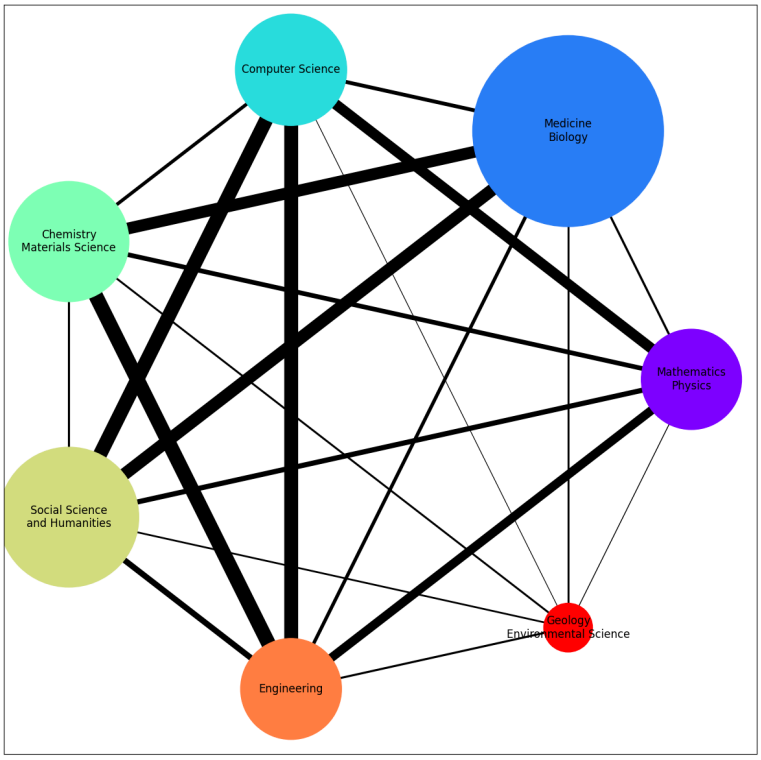
Louvain Communities for Year 2009



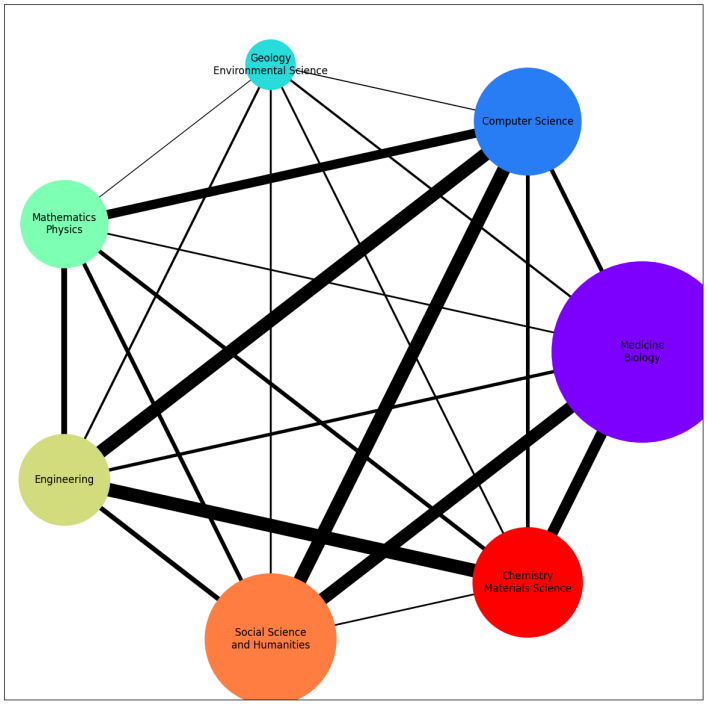
Louvain Communities for Year 2010



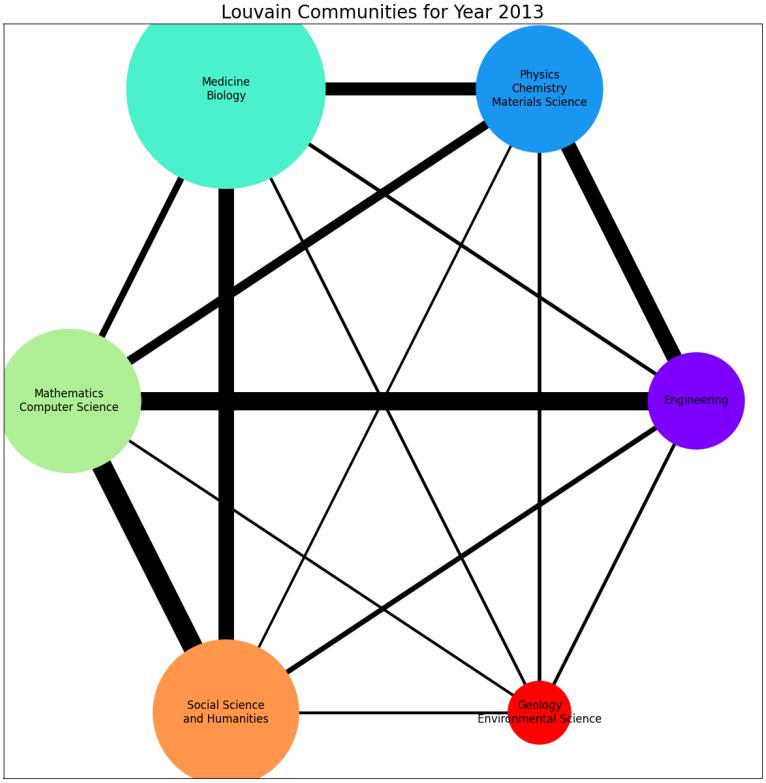
Louvain Communities for Year 2011



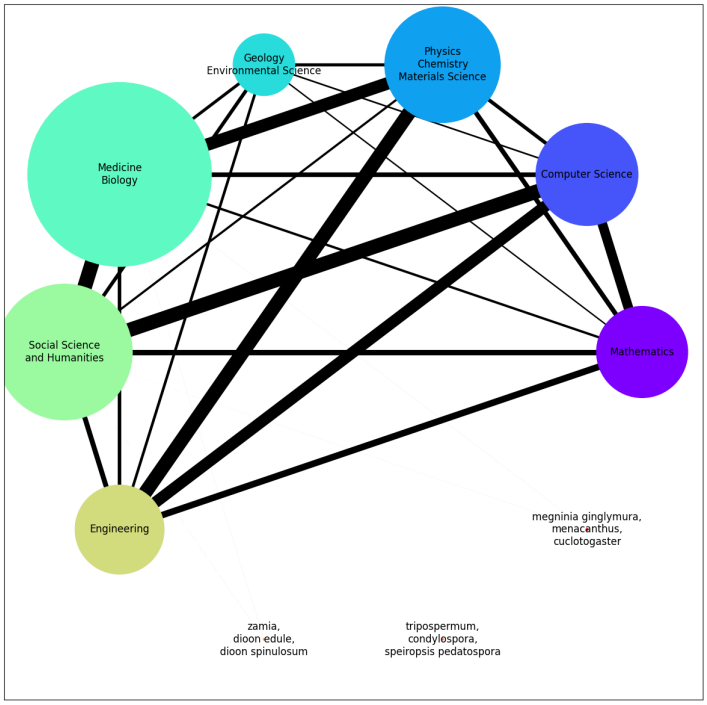
Louvain Communities for Year 2012



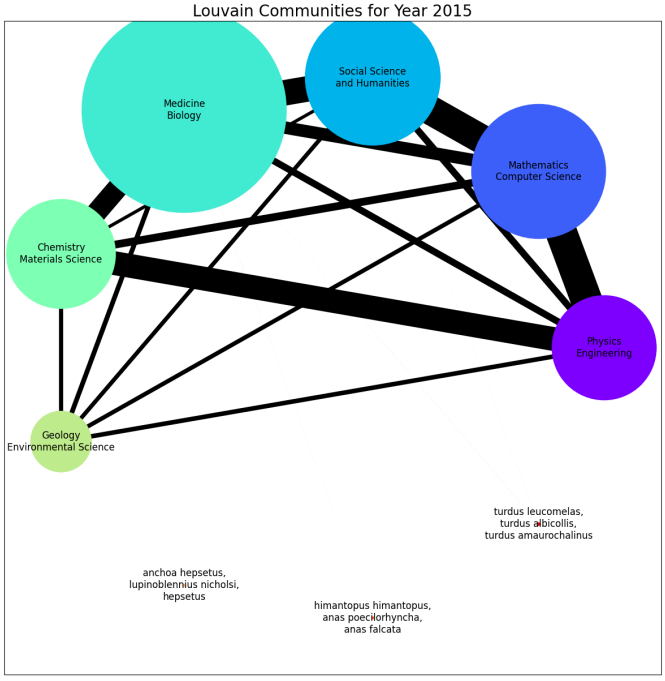
Louvain Communities for Year 2013



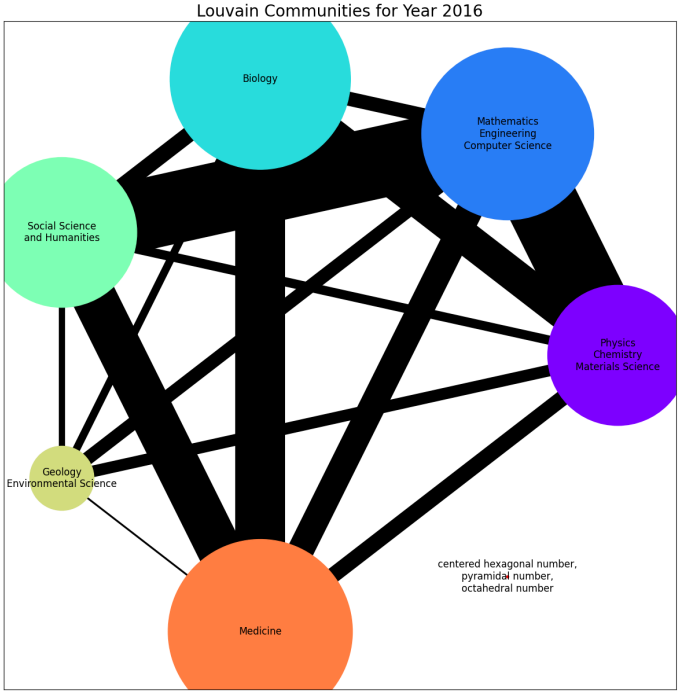
Louvain Communities for Year 2014



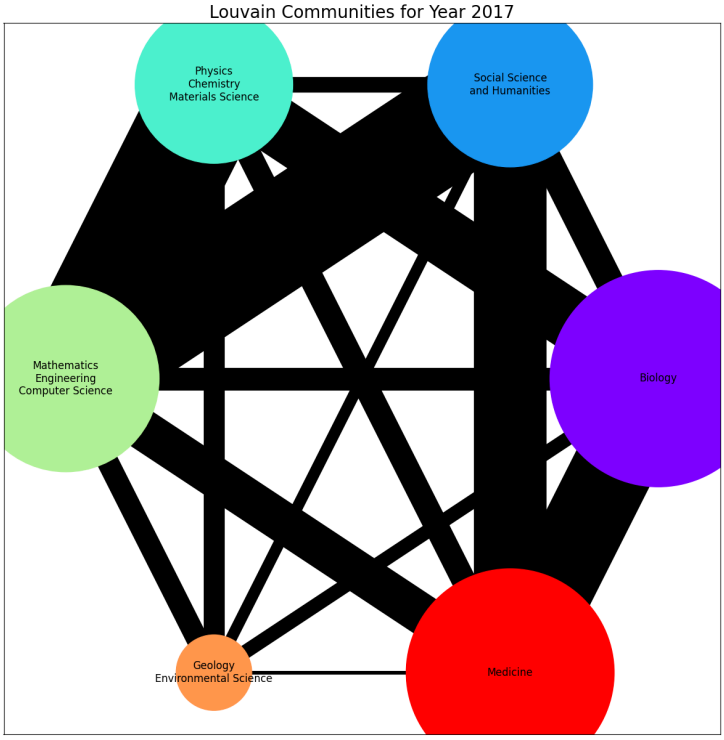
Louvain Communities for Year 2015



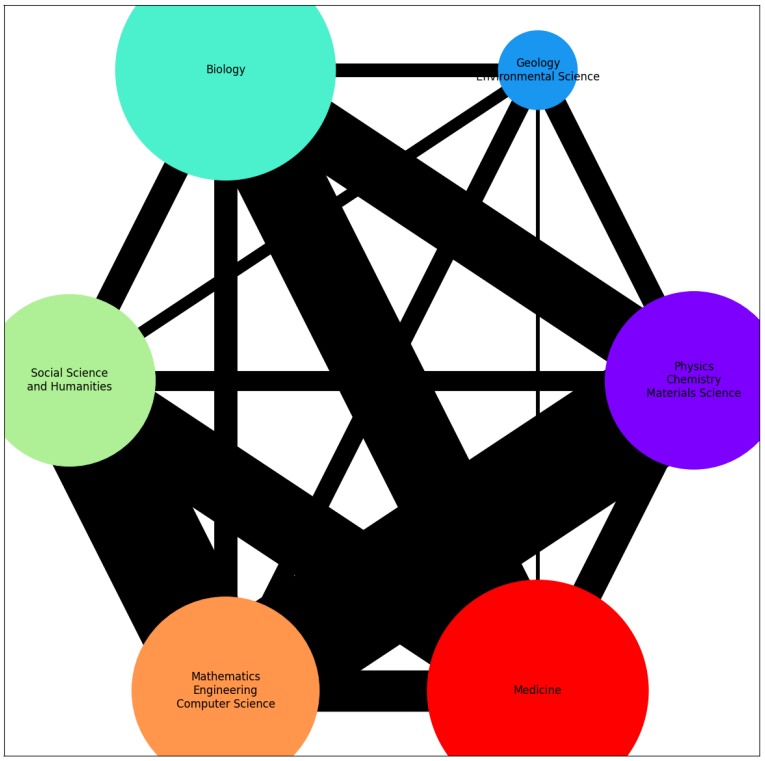
Louvain Communities for Year 2016



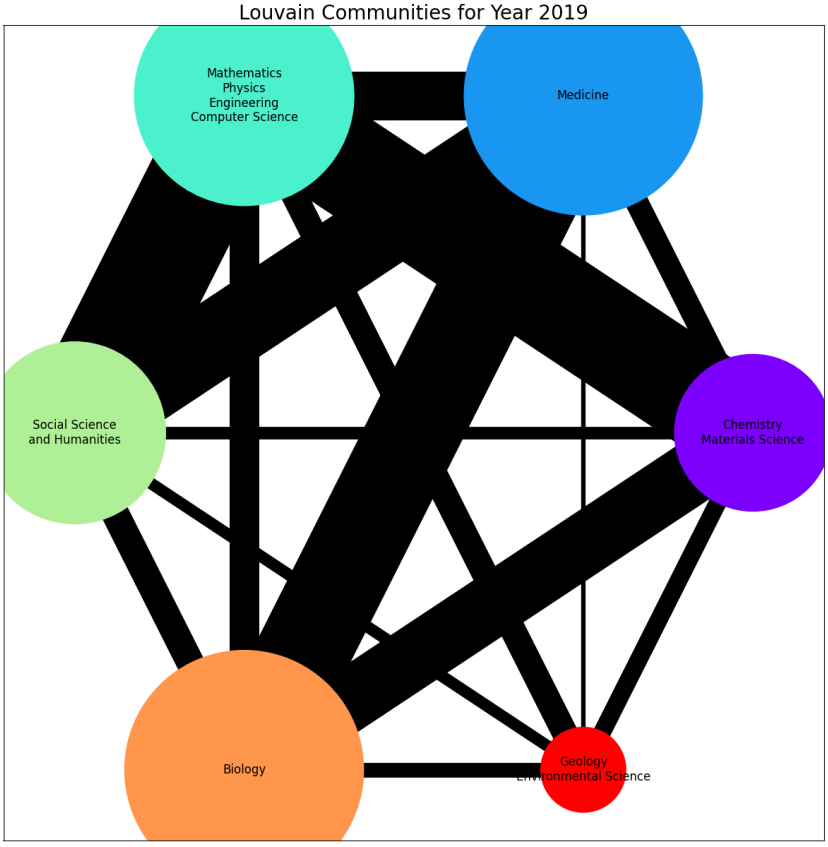
Louvain Communities for Year 2017



Louvain Communities for Year 2018



Louvain Communities for Year 2019



The Louvain community detection results across the years reveal various collaboration patterns and how different scientific disciplines interact.

- **2000-2003**: The communities are relatively distinct, with clear separations between fields like Medicine/Biology, Chemistry/Materials Science, and Mathematics/Physics. Collaborations between these fields begin to strengthen, especially between Medicine/Biology and other fields like Chemistry and Engineering. Early signs of collaboration between Computer Science and Social Sciences also start to emerge.
- **2004-2007**: Interdisciplinary connections grow, with increasing interaction across fields. Physics joins the Chemistry/Materials Science community, while Mathematics remains in a separate cluster. Strong collaborations continue between Medicine/Biology and Chemistry/Materials Science. The thicker edges representing stronger connections indicate significant interdisciplinary work, particularly between Engineering, Physics, and Computer Science.
- **2008-2011**: During these years, Medicine/Biology emerges as a dominant community, with increasing collaborations across other disciplines. Mathematics and Physics reunite to form a single community. Social Sciences and Humanities begin to collaborate more closely with technical fields like Mathematics/Physics and Computer Science, highlighting the growing interdisciplinary nature of research during this period.
- **2012-2015**: Physics, Chemistry, and Materials Science form a strong, interconnected community. Social Sciences also show stronger ties to technical fields. In 2013 and 2015, Mathematics was closely linked with Computer Science, while in 2014, it separated into its own community. Similarly, in 2013-2014, Physics joins Chemistry/Materials Science, but in 2015, it aligns with Engineering.
- **2016-2019**: Communities continue to consolidate. Medicine and Biology separate into distinct communities, with increasingly robust collaborations with Mathematics, Computer Science, Physics, and Chemistry. By 2019, Physics merges with the Mathematics/Engineering/Computer Science community, highlighting stronger interdisciplinary links. Social Sciences and Humanities maintain close ties with technical disciplines, reinforcing the growing importance of interdisciplinary research. Across all fields, previously isolated scientific disciplines show greater collaboration, indicating a trend toward more integrated, interdisciplinary work in the later years.

Results

Our analysis revealed that:

1. **Medicine and Computer Science Lead**: As the graphs indicate, Medicine remains the most researched field, with nearly 1.75 million publications. Computer Science showed a significant increase in research output, reflecting the growing demand for technological research in the modern era.
2. **Interdisciplinary Growth**: The Louvain community detection algorithm uncovered growing collaborations between disciplines such as Computer Science, Biology/Medicine, Chemistry, and Social Sciences. Over the years, the strength of these interdisciplinary ties has increased,

especially from 2016 onwards, highlighting the trend toward more integrated, cross-disciplinary research.

3. **Clustering Analysis**: K-means clustering revealed that reducing the number of clusters to 12 helped separate research fields like Biology and Medicine into distinct, coherent groups. However, at lower cluster counts (e.g., K=10), some unrelated fields like Psychology and Mathematics were grouped together, suggesting that a higher number of clusters is necessary to maintain meaningful distinctions between disciplines.
4. **Community Detection**: The results of Louvain community detection revealed the formation of several dominant research communities over the years. These visualizations show clear separations between fields like Medicine/Biology and Chemistry/Materials Science, while technical fields like Computer Science, Physics, and Mathematics formed their own robust communities in the later years. These findings underscore the increasingly interdisciplinary nature of academic research.

Impediments

Several challenges were encountered during the evaluation phase:

1. **Large Data Processing**: Handling the massive dataset of 12 million rows required significant computational power and time. To address this, we distributed the workload across multiple machines using parallel computing and optimized performance by using a pickle file, which significantly reduced load times for further processing.
2. **Memory Management**: Processing the 10GB files in smaller chunks was necessary to prevent memory overflow and maintain performance. This approach ensured that operations could be executed efficiently without running out of memory.
3. **Data Inconsistencies**: Some papers had incomplete metadata, particularly regarding author or organization information. We filtered out incomplete rows while ensuring that the final dataset remained representative of the research trends.

Future Work

There are several avenues for future exploration, including:

1. **Advanced Prediction Models**: Developing machine learning models, such as neural networks or ensemble methods, to predict emerging research fields and identify potential "hot topics" in academic research.
2. **Enhanced Network Analysis**: Building on the current community detection results, we could explore more advanced network analysis techniques to quantify the strength of interdisciplinary ties, investigate influence metrics (like betweenness centrality), and identify key researchers driving collaboration between fields.
3. **Expanding the Timeframe**: Extending the dataset beyond 2019 would allow us to capture recent trends and investigate the impact of major global events, such as the COVID-19 pandemic, on academic research output and interdisciplinary collaborations. Note that the new data is in a different format and was not found on the original website from which we downloaded the data.

Conclusion

Our project successfully analyzed academic research trends using data from the Microsoft Academic Graph. We identified dominant fields, such as Medicine and Computer Science, and uncovered significant interdisciplinary collaborations across various research areas.

By applying TF-IDF, K-means clustering, and community detection, we visualized the evolving dynamics of research fields and collaborations. These insights highlight the growing importance of interdisciplinary work and provide a foundation for future analysis, such as prediction models and extended network analysis.