**Speech Research Trend Visualization Report**

Xingyu Shen, Yueqian Lin, Zhixian Zhang

<https://github.com/linyueqian/Speech-research-trend-vis>

**I. Introduction**

This final project report aims to visualize the research papers related to speech in a clear and concise manner, providing insights into the development and trends in this field. With the rapid development of deep learning in speech, it is essential to stay current on the most recent trends.

Our visualization goals include exploring how different authors collaborate with others and how deep learning-related knowledge intertwines with other disciplines such as physics or medicine, recognizing trends in the growth and development of intelligent speech, and comprehending how these trends change over time from the abstract of academic papers. We also aim to examine the correlation between citation numbers and publication numbers and to uncover patterns in the paper's publication time. These patterns can help researchers gain insights into common publishing practices and the selected categories of research papers.

The research questions of this visualization project are to explore the connection between different authors and between deep learning and knowledge in other fields, identify the frequent word trends of the speech-related papers, examine the correlation between citation numbers and publication numbers, and identify whether authors tend to select specific publication times for their papers and what patterns we can identify in these practices. By answering these questions, we can gain a better understanding of the most recent advancements in intelligent speech, and their influence on the field.

**II. Method**

1. **Data collection, cleaning, and analysis**

We obtained the data from academic papers on a public scientific exploration and discovery tool called “Semantic Scholars.” It provides details of over 200 million academic papers sourced from various publishers, data providers, and web crawls. Using the Python package of semantic scholars, we used the query of “speech” to obtain 10,000 papers each year from 2020 up to now, according to the first 10,000 results given by the search of the query, making the total dataset more than 32,000 entries. The features we get from each paper include title, author names, author ids, publication date, reference count, citation count, influential citation count, venue, and field of study. By the list of author ids, we get detailed information about every author, including their own paper count and total citation count. We only included the data of the first author of every paper for concision. We also calculate the rate between publication count and citation count, and level authors from one star to three stars according to the mean and variance of publication\_citation rate. Since the raw data crawled are mostly strings, we cleaned them into lists, numbers, time data, etc. The title, authors, and abstract are nominal data. The venue and field of study are categorical data. The publication date, 1st author paper count, 1st author citation count, paper citation count, publication\_citation rate, and reference count are numerical data. 1st author's star is ordinal. In each visualization, the data will be further cleaned to adapt to the visualization tool.

1. **Data visualization**

The interactive network graph in the network section displays academic collaborations between authors. To provide a clearer view of the network, only authors whose papers are in the top 100 citation numbers papers are selected. Using Python, different combinations of these authors with various co-authors are generated, resulting in a .gexf file that can be imported into Gephi software. Gephi is an open-source software that visualizes and analyzes networks and graphs. Using the Fruchterman Reingold algorithm, the undirected relationships between different authors are grouped. Nodes represent authors, and edges represent collaborations, with size and color indicating weighted degree. The color of each node corresponds to its weighted degree. This visualization, built using the Sigma.js library, allows users to analyze the network structure, identify key authors and communities, and gain insights into their relationships.

To provide a deeper understanding of the larger nodes, we manually select six influential authors and scrape their entire body of academic work. A separate network is generated from this newly scraped data, showcasing the collaborations between the influential authors and their collaborators. The network is processed with a force analysis algorithm.

The interactive word cloud section shows the Top 250 common words in the abstract part of the papers. To clean the dataset, Python is used to filter out words in a stopword list and does the word count. Echarts-wordcloud package is applied to make the words into a word cloud and adjust its shape to a microphone according to the provided mask image.

To better understand the change in topics over time, the bar chart race is created to dynamically display the changes in word usage over time in the abstract part. Python is used to filter out some stop words and create a new CSV file, which contains the date and the top ten cumulative occurrences of the words and their corresponding counts on that day. We also manually exclude some words that are not helpful in analyzing the trend. Then, some adaptations are made to the barchartrace package to create the bar chart race.

The chord diagram shows the interdisciplinary correlations between different fields of the papers. The papers that are mono-field studies are ignored since they have no relationship with interdisciplinary studies. We constructed the matrix for the chord diagram based on the number of co-occurrences of two fields and which is the major field. The color of the ribbons depends on the number of major fields in the interdisciplinary course. We use Python to clean the data and construct the matrix while using d3.js to make the interactive chord diagram.

The calendar heatmap shows the relative number of papers published each day. We use Python to group and count the number of papers published every day, then use d3.js to draw the calendar and the colors. We change the date into date data type in d3, and we can see the date and the number of papers published that day by hovering over the block. The more papers published, the darker red is in the block.

The quadrant chart is created to show the relationship between the publication number and citation number of 1st author for 4500 papers collected. Python is used to process the dataset and generate the JSON file containing the author name, paper count, and citation count. The logarithm is applied to make the data fit the chart. Then, some adaptations are made to the ScatterPlot package to create the quadrant chart.

1. **Interactive website building**

We have made our visualizations easily accessible by publishing the entire website on Github Pages, using HTML, CSS, and JS files. To simplify the process of merging and comparing different CSS files used on each page, we've embedded all separate visualizations in the portfolio using the <iframe> element. Additionally, we've carefully selected colors and font styles for the background, navigation bar, and information panel to maintain a cohesive design language throughout the website. Although the interactive visualizations are primarily designed for PC users, we've ensured that the website is responsive, with adjustable navigation bars and self-adaptable visualizations, allowing users to access the page on their mobile devices. The website can be found at [vis.yueqianlin.com](http://vis.yueqianlin.com).

**III. Results**

We have created five visualizations that depict different aspects of our research questions. Each visualization is interactive and provides a clear representation of the data. We provide explanations of each visualization and how it relates to our research questions.

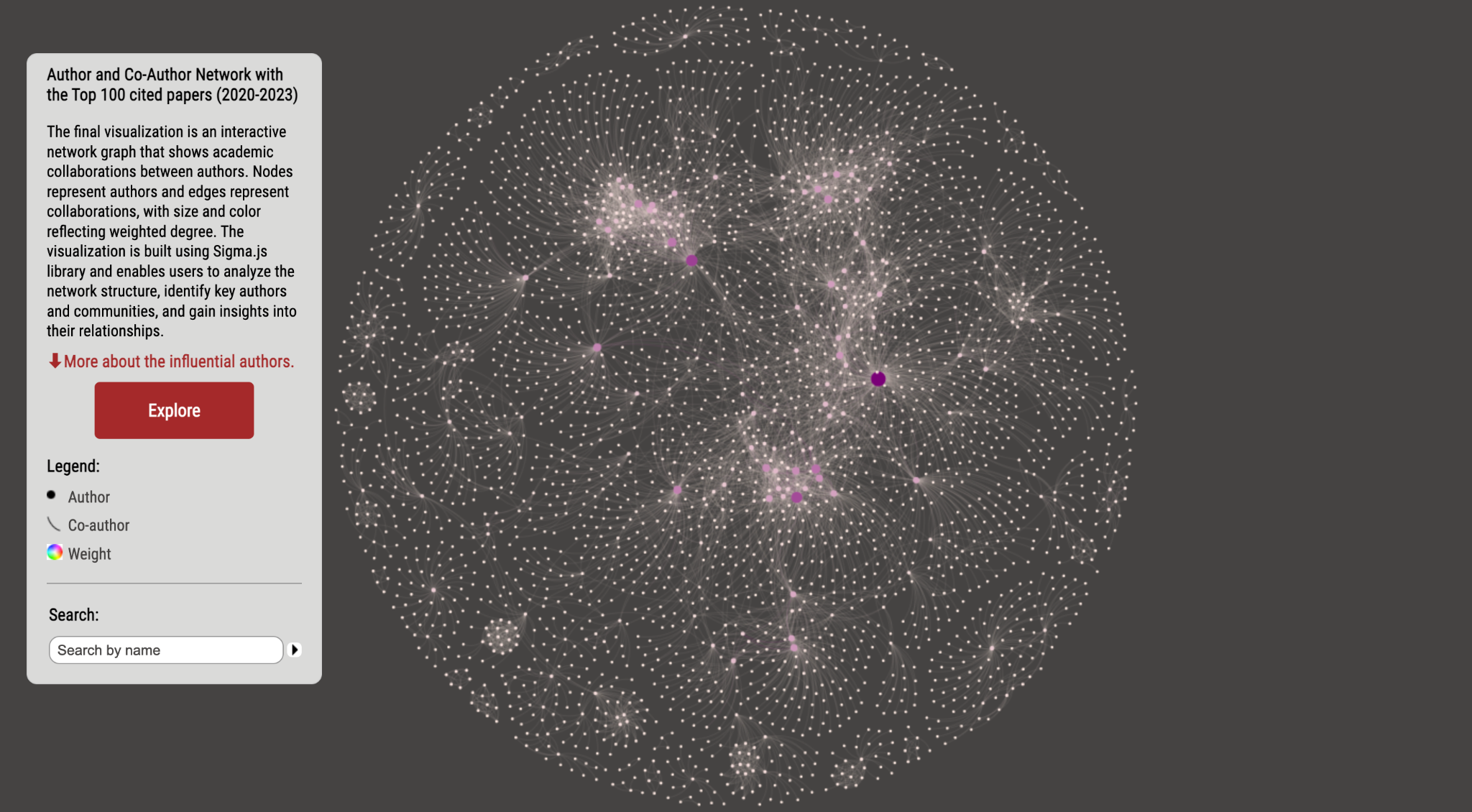


Figure 1: overall co-author network

The author and co-author network provides an overview of the groups and communities in the speech research area over the past three years. The center of the network displays brighter edges and closer connections between authors, while the outer circle is more sparse but still contains some smaller groups. Users can zoom in/out of the graph to examine specific regions and perform dynamic queries by inputting an author's name in the search bar. Based on the node size and color, influential authors in the research area include Junyu Li, Shinji Watanabe, Tara N. Sainath, Yu Zhang, and Zhuo Chen.

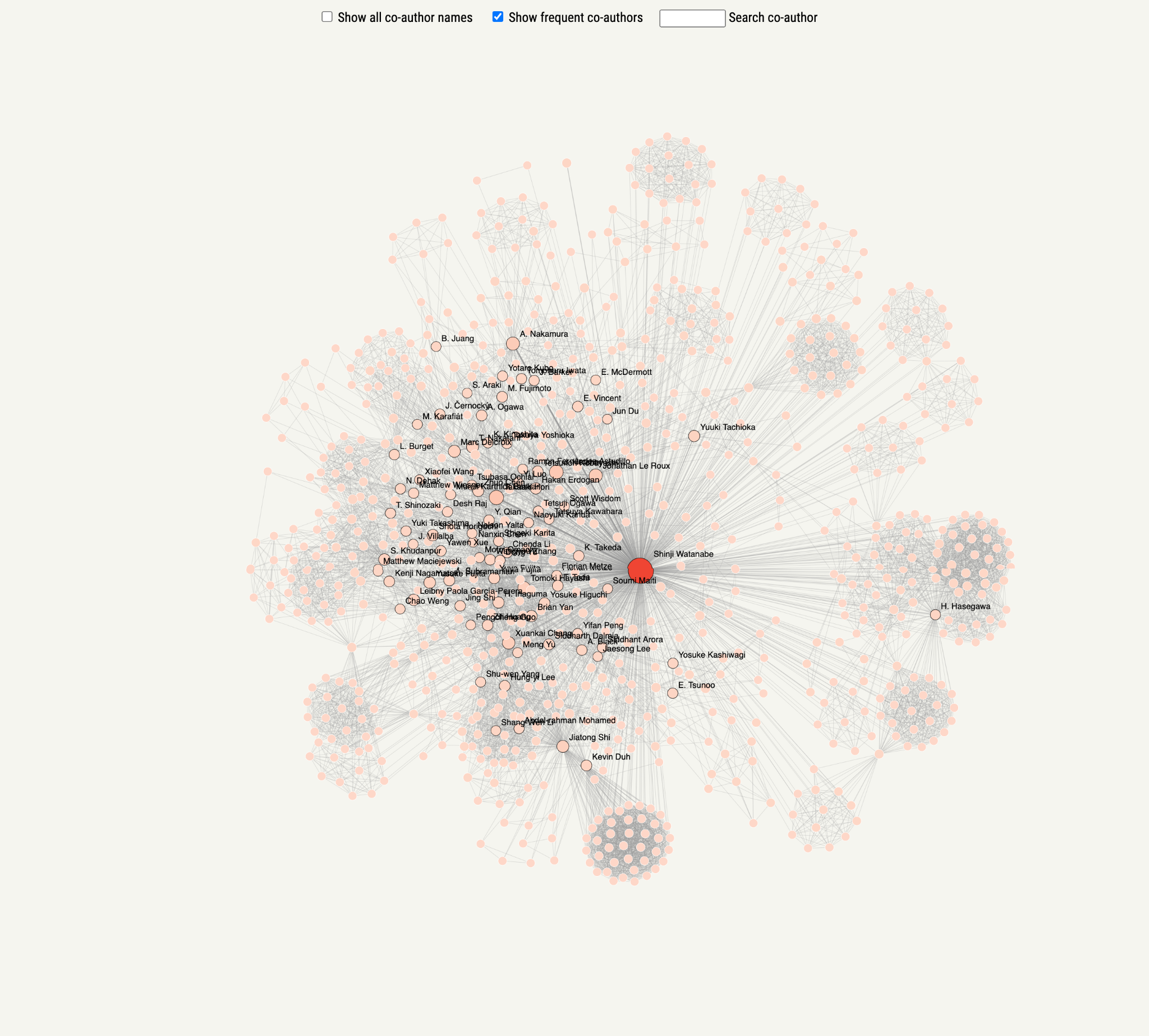


Figure 2: individual co-author network

One of the individual co-authorship networks is shown in Figure 2. In this example, Professor Watanabe has numerous collaborators, and many of them appear more than once. Interestingly, there are internal groups for different co-authors, possibly due to research groups led by the same professor. Users can drag a specific node to understand a local region better and check the “show-all co-author names” or directly search the co-author for a more detailed understanding of the co-authorship network.

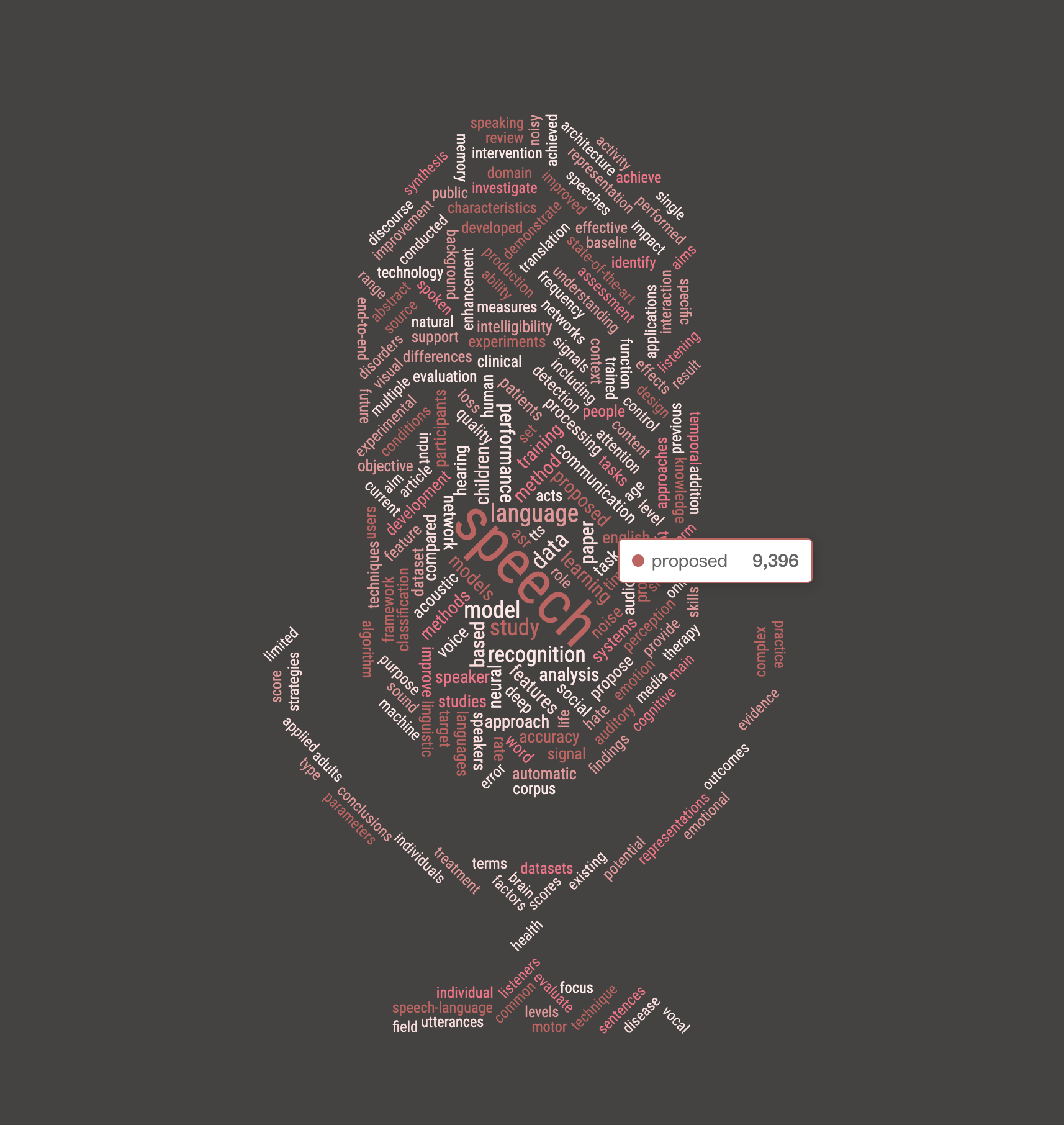


Figure 3: interactive microphone-shaped word cloud

In the microphone-shaped word cloud, the size of each word reflects the number of occurrences in the abstract part. And detailed data will show up when hovering over the words. According to the word cloud, most of the frequent words are closely related to speech research, such as speech, recognition, and language.

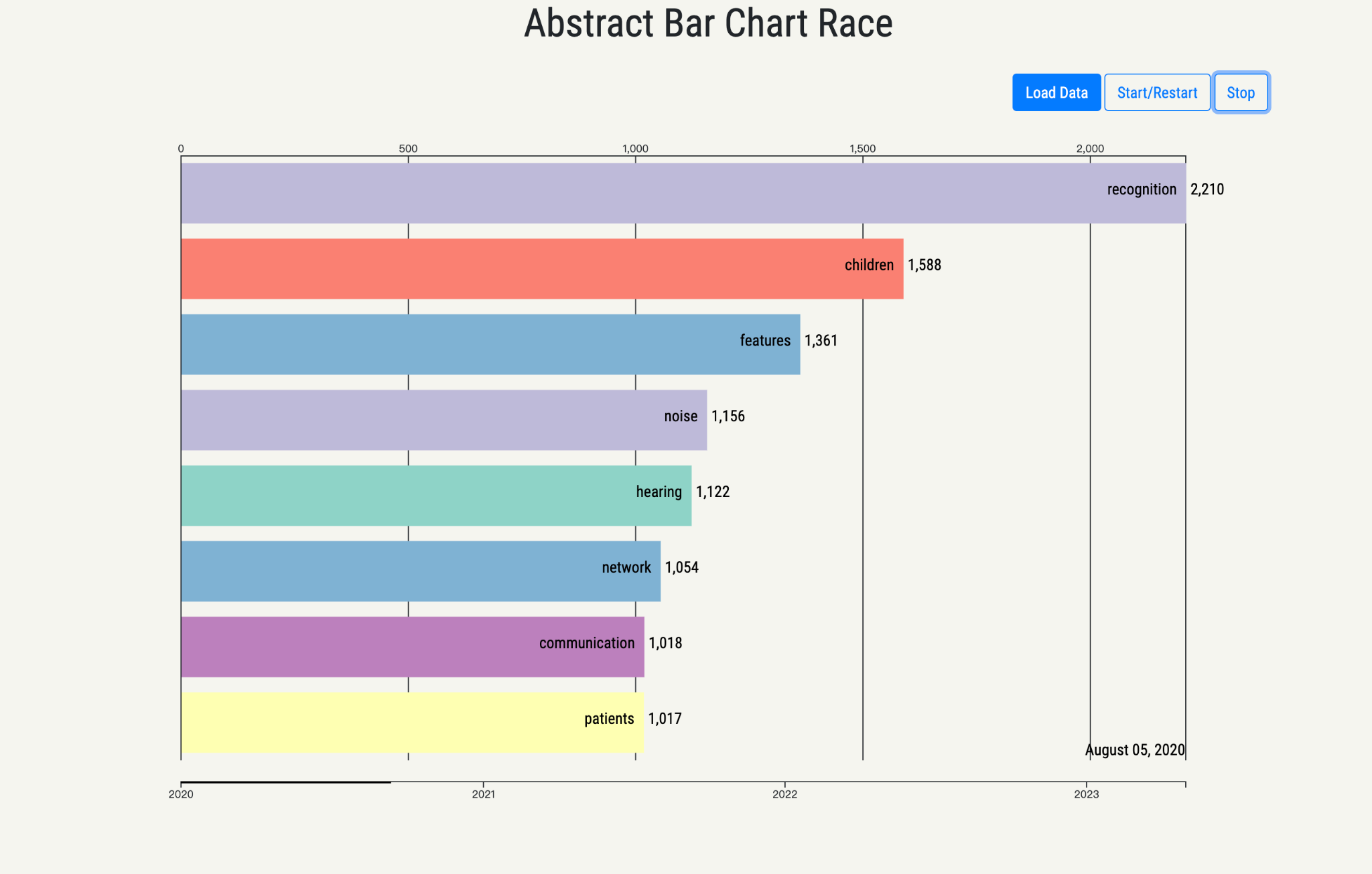


Figure 4: bar chart race for occurrence of words in paper abstract part

In the bar chart race, the length of each bar reflects the number of occurrences. The bar chart race is generated after pressing the 'Load Data' and 'Start/Restart' buttons and the animation shows the cumulative most frequent eight words changing over time. The 'Stop' button can be pressed to stop the animation and see the specific data at that time. As well as a few words like 'recognition' and 'features', which are always at the top of the list, we were surprised to find that 'children' and 'patient' are also frequently used in the abstracts of speech recognition papers.

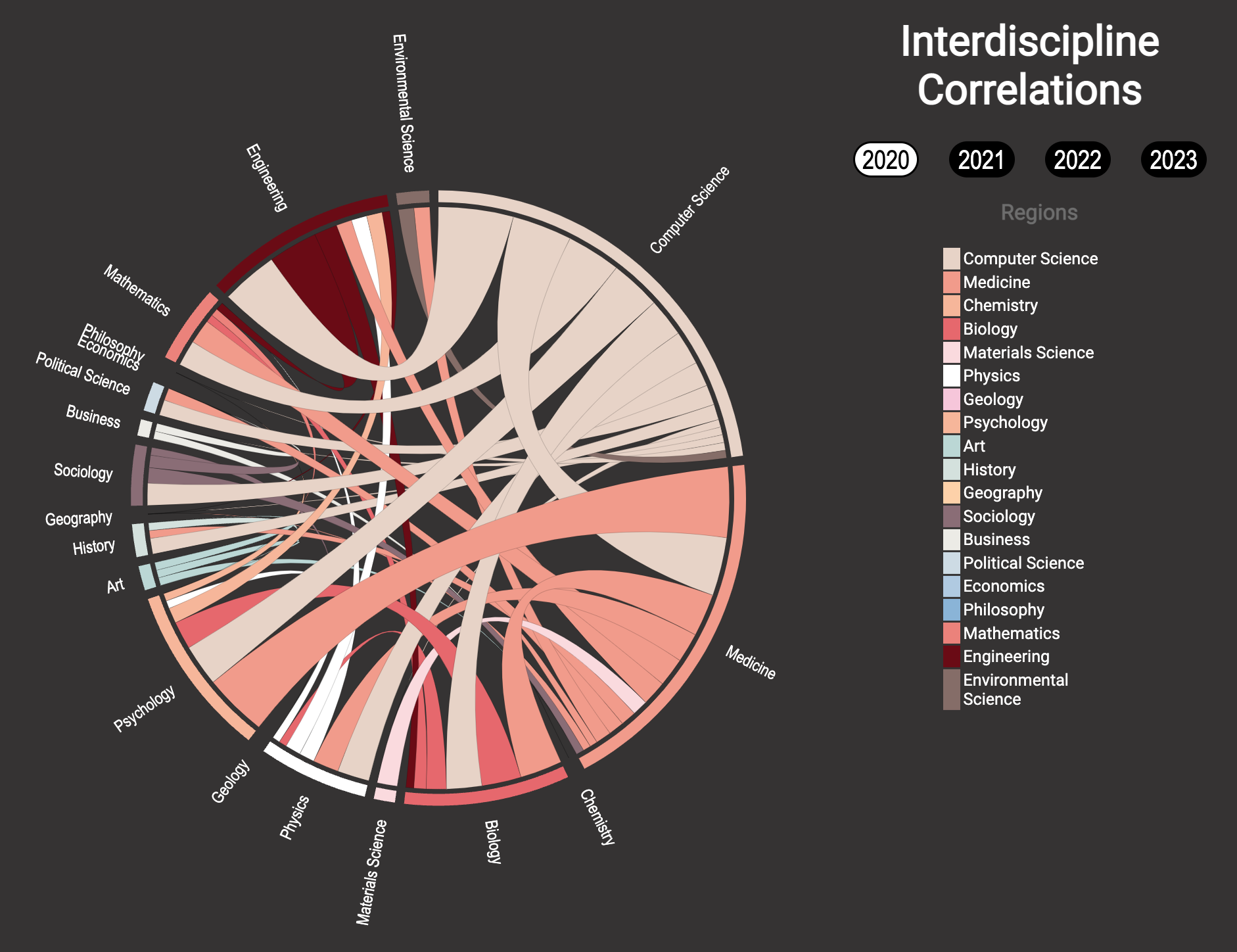


Figure 5: Chord diagram for interdisciplinary correlations

Since the color of the ribbon in the chord diagram represents the major field in the interdisciplinary studies, we can see if a field mainly acts as a major field or minor field. As shown in the diagram, in 2020, computer science and medicine mainly acts as major fields in cross-field studies. Engineering, psychology, biology, mathematics and physics mainly act as minor fields. It has the correlation that basic subjects (like math, physics and biology) serve as minor fields in interdisciplinary studies with integrated subjects like computer science and medicine. However, when we click to other years, we can see that this trend may gradually change. Also, from 2020-2023, the proportion of computer science has become larger and larger. We can see the increasing attention paid to it.

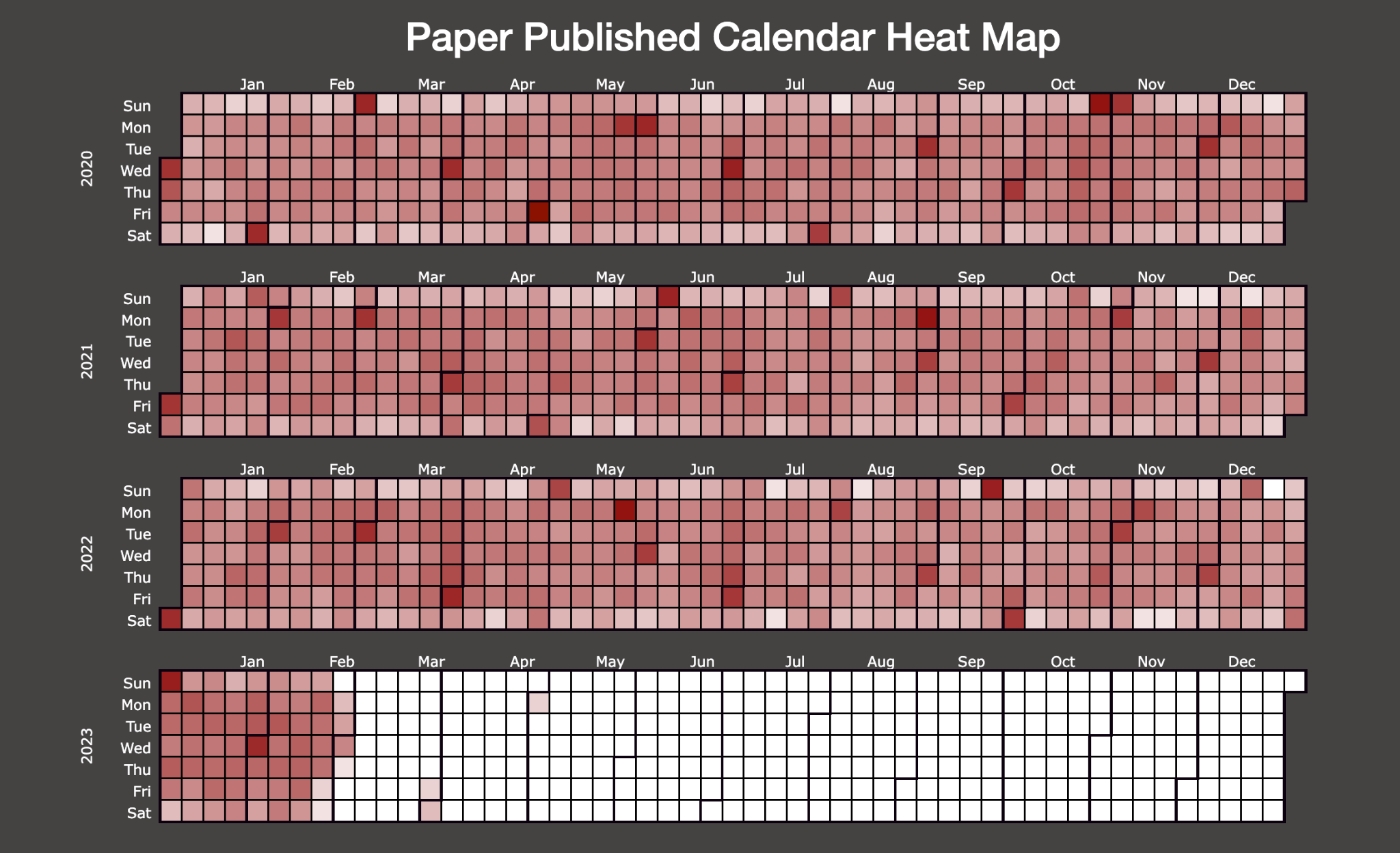


Figure 6: quadrant chart for publication count and citation count

The heatmap shows the publication's popularity throughout the past three years. We can see that the middle five columns are generally darker than the top and bottom rows. It means that researchers are more likely to publish papers on weekdays, instead of weekends. Another trend is that every first day of the month is the publication climax. We can see that nearly every month has a deep red block on the first day. We suspect that it is because some monthly journals make the publication date to be the first day of the month.

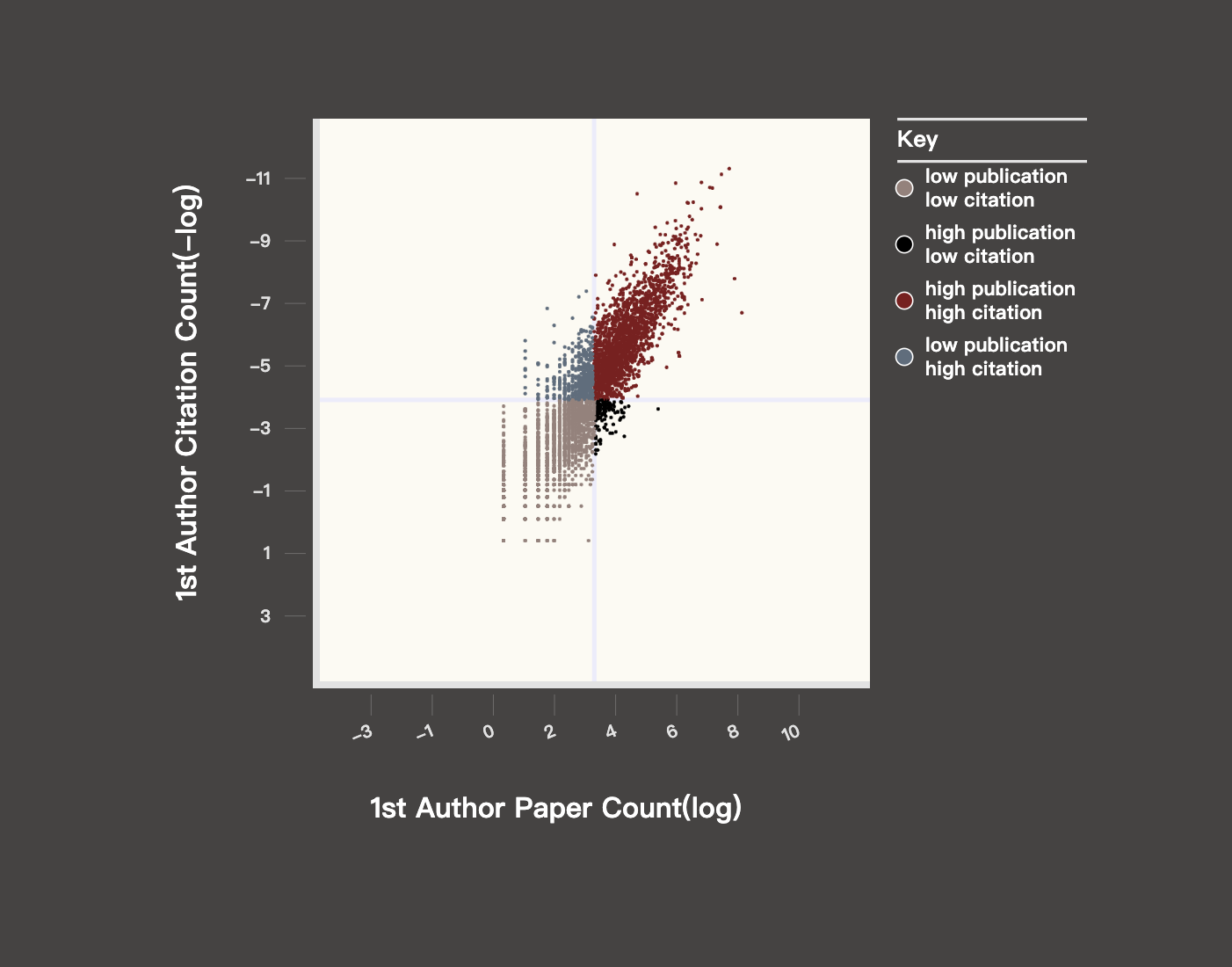


Figure 7: quadrant chart for publication count and citation count

In the quadrant graph, the dots are divided into four parts based on averages and are coloured differently. By clicking on the dots, the graph will zoom in and the selected dot will be placed in the centre. Detailed information such as the author's name, publication number and citation number is also displayed next to the dot. Most of the dots are linearly distributed, which corresponds to our instinct.

**IV. Discussion**

We use the what-why-how analysis framework to evaluate the visualizations regarding the user task, interaction and efficiency.

What: The user task of this final project report is to create a visualization that provides insights into the development and trends of speech-related research papers. The visualization aims to explore the collaboration between different authors, and the relationship between deep learning and other fields, identify trends in speech-related research papers, examine the correlation between citation numbers and publication numbers, and uncover patterns in the publication time of papers.

Why: The goal of the visualization is to help users gain insights into the field of speech-related research and identify patterns and trends that may not be apparent from reading individual papers or abstracts. The visualization aims to make it easier for users to explore the collaboration between different authors and identify interdisciplinary connections, track the evolution of research topics and identify emerging areas of interest, and identify relationships between citation numbers and publication numbers.

How: The visualization effectively demonstrates the connections between authors, trends of words in abstracts over time, the correlation between the paper number and citation count, publication date heatmap, and chord diagrams of various disciplines. Users can make dynamic queries to search for specific authors on the network diagram, hover over words to see detailed counts, select specific years for a clearer view of interdisciplinary correlations, toggle specific years to zoom in on the heatmap, and click on dots in the quadrant chart to display detailed information.

Future improvement: There is still room for improvement in the data processing part. Some author names may be duplicated due to abbreviations, making them difficult to identify. Additionally, the Semantic Scholar API has query frequency limitations, making it challenging to deploy visualizations dynamically for a specific keyword. The stop-word list used for the word cloud and bar chart race is not tailored for speech papers, leading to irrelevant words in the analysis. Furthermore, for more accurate analysis, each word should be counted only once instead of counting all occurrences in the abstract. The heatmap can be paginated by different conferences, making the comparison between venues easier. The quadrant chart could benefit from additional attributes such as author age, and for privacy concerns, author names can be hidden, displaying only the research area. Furthermore, conducting a more detailed analysis of the outliers in the quadrant chart could lead to better insights.

**V. Conclusion**

Our visualizations have been instrumental in generating knowledge and providing visual support for our research questions. Our results show that authors are grouped into different communities and collaborate more within their communities. Furthermore, the field of speech recognition remains the most researched field across the community, while the proportion of computer science disciplines gradually increased over the years. In addition, we observed that authors tend to publish their papers at the beginning and end of the month.

**VI. Roles and responsibilities of each team member**

ZZ was responsible for collecting and filtering the total dataset, cleaning the data, and creating visualizations of the chord diagram and heatmap. YL collected and filtered the author dataset, cleaned the data, created visualizations of the network, built the skeleton of the demo website, and worked on the poster. XS filtered and cleaned the data, created visualizations of the word cloud, bar chart race, and quadrant chart, and worked on the poster. Together, we used our individual skills and expertise to contribute to the project’s success.