CS6630 Project
Visualizing CS Collaborations

Process Book

Youjia Zhou, Wenzheng Tao, Mingxuan Han

11/09/2018

Contents

1 Member Information

Member 1: Youjia Zhou, zhouyj96180@gmail.com, u1208920
Member 2: Wenzheng Tao, wenzheng.tao@utah.edu, u1210098
Member 3: Mingxuan Han, u1209601@umail.utah.edu, u1209601

2 GitHub Repository

https://github.com/zhou325/dataviscourse-pr-VisCsCollaborations

3 Overview and Motivation

The DBLP Computer Science Bibliography dataset contains more than 1.2 million bibliographic records. Hence, for researchers, it is a useful tool to trace the academic works and to get bibliographic details when composing the list of references for the new papers.
While there are a lot of works on cs rankings based on DBLP dataset, we found that it is also interesting to study the collaborations in computer science between universities using this dataset. The motivation of this project is from a small talk. One of our group member would like to know which university has the most relationships with other institutions in computer science, indicating the willingness of this university to communicate with other institutions over the world. **Based on the DBLP dataset, we define such relationship as the number of publications two institutions have worked with together.** And we believe that we are not only people who are interested in such kind of questions, like which university are the most active in computer science and which institution they should go if they are interested in doing research in certain specific areas in computer science. It is become a natural choice for us to run this project.
Overall, our project is supposed to provide the interactive visualization of worldwide affiliations research achievements in top CS conferences. Publication amount as well as research cooperation will be displayed together. We will use group of charts to show the relationships among the universities in the dimensions of location, topic field, and time. According to the data scales, the charts can be divided into three groups: world view, contrast view and certain university view.

3.1 World View

Firstly, we put every university on a world map so that it will be convenient to watch and discover the geographical distribution of research affiliations. Users can select the period and universities of interest, and the corresponding data will be displayed. We link universities that have collaborated in CS publications so that users will be able to see the development of research affiliations among selected universities and how this relationship changes over time. Meanwhile, users can recognize how many cooperative companies one university owns and among its collaborators who is the closest one. In addition, we also aims to visualize the overview correlations between the CS rankings, closeness geometrically and cooperation variety in this part.

3.2 Contrast view

Our visualization will help users look at the differences between selected universities in CS research achievements, mainly through the contrast-view group of charts. It allows the user select the universities by clicking on the nodes or other methods, such as drawing a rectangle to include the nodes, which is helpful to build up a contrast among universities in a certain area. We will apply bar chart to compare the total publication amount in four sub fields of CS, those are AI, System, Theory and Interdisciplinary Areas. And there will be a parallel ranking table, where the user could have a precious overview of what are the ranks of the universities in total and the subfields. Also, the user could see and resort a table through clicking on different domain to see the exact ranking of the universities.

3.3 Certain university view

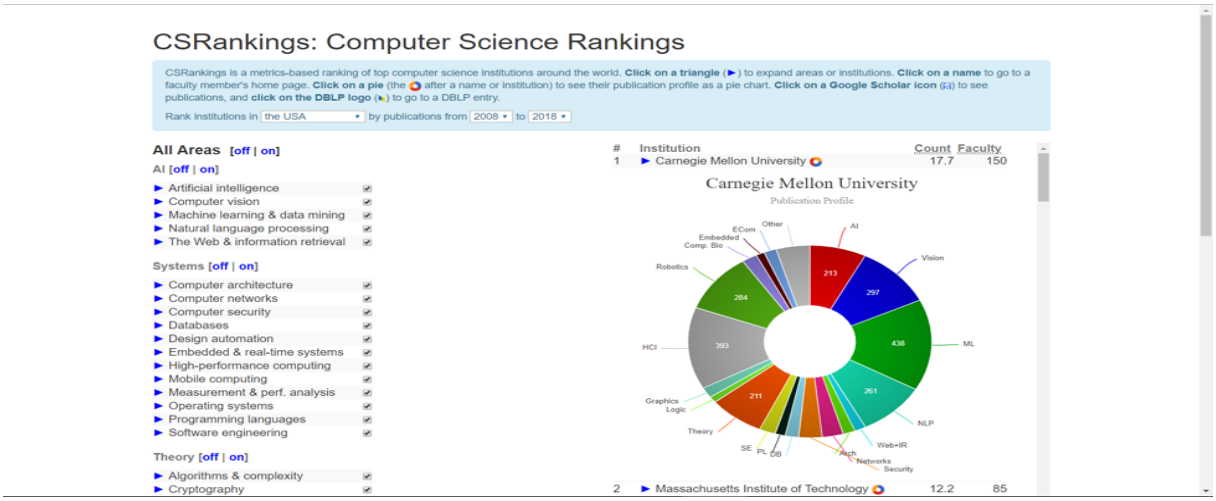
If the user is interested in learning more information of a certain university, our visualization allows the user selecting it and see the information such as university name, CS ranking, fields of research, number of cooperative universities and geographical parameters in a note panel. By click the certain field, the user will be able to see the ratio of the corresponding conferences in a donut chart.

4 Related Work

The topic of this work is scholar data visualization. There are several works that provided us with great inspiration.

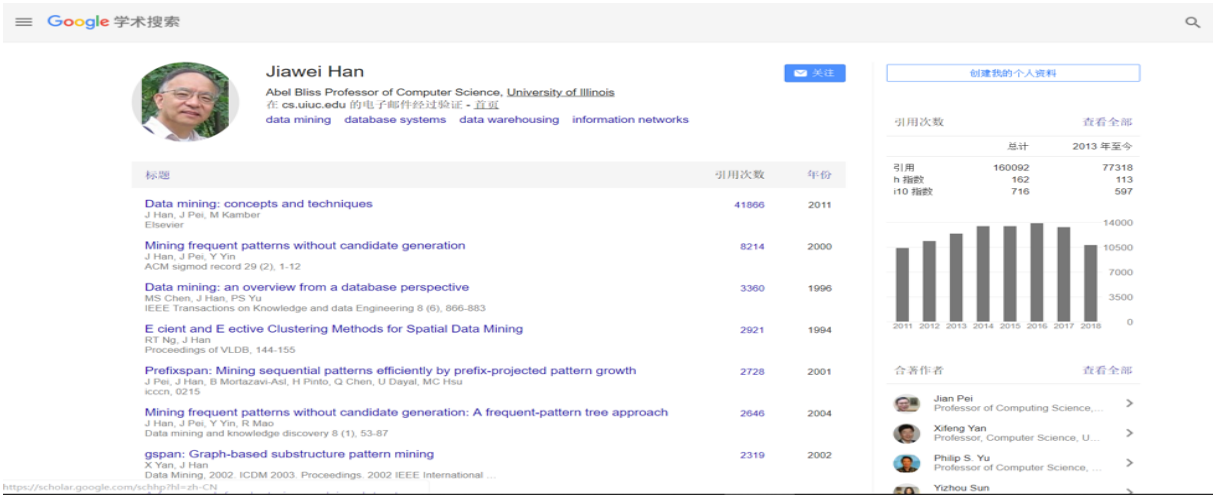
4.1 Csrankings.org

- CSRankings is a metrics-based ranking of top computer science institutions around the world. As stated on their main page, their ranking is designed to identify institutions and faculty actively engaged in research across a number of areas of computer science, based on the number of publications by faculty that have appeared at the most selective conferences in each area of computer science.
- They set up the project by classified the top conferences into 4 different areas and sorting up the data from DBLP, a famed CS publication record website. This way it displays the area composition of these institutions.
- Due to the reputation of the selected conference and the trustworthiness of the DBLP, its ranking, which is calculated by the geometric mean count of papers published across the areas, quickly gained it reputation against the ranking given be other affiliations, such as US-News.
- And it collected the webpages of the authors so that the user could be directed to the authors homepage conveniently. It applied pie chart to display the percentage of the publication profile.
- It also provides the users much freedom in choosing the indicators to make their own rankings. The indicators include the interval of publication date, the regions of the institutions, the areas and more specifically, the conferences of the publication.



4.2 Google scholar

- Google scholar is a commonly used tool in nowadays. As stated by itself, it is a web search engine that could search articles online with no charge and its index include most of the worldwide publication.
- Researchers tend to build up their personal page on google scholar. It is now a popular method for the interested to learn about the research records of scholars. And google scholar provide the co-authors information as well, so the users could easily have a glance at the co-author network.
- Google scholar has large number but somewhat noisy data, it did some process about the data and the generated indicator such as citation and h-index, which are both respected methods to do scholar evaluation. However, the user might want more visualization techniques to help depict the scholar records more concretely, rather than simply using the numbers.

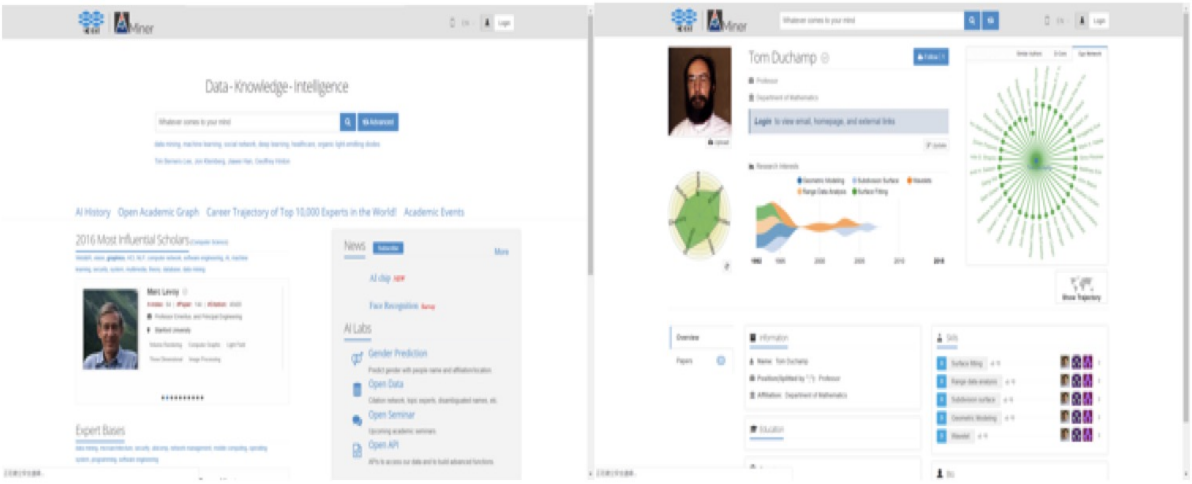


4.3 acemap.info

- Acemap is a project from Shanghai Jiaotong Univ, aiming to apply visualization to display the scholar records. As is stated on its main page, the goal is to make academia visible.
- It did analysis to affiliations, authors, and topics. It built up specific page for each individual researcher, affiliation and topic to help user know about their info. Acemap recently published several statists about the top conferences, including translating the outline of the papers, rank the authors according to the publication frequencies, and topic evolution.
- Acemap can be recognized by present the network relation in force-directed map, which is pretty similar to the formation of the universal galaxy.

4.4 aminer.org

- Aminer is abbreviated from academia-miner, it is also a scholar project using visualization techniques, mainly displaying records in computer science. It is from Tsinghua university.
- It set up personal webpage for each researcher by automatically acquire data from the researchers real home page. Therefore, the project could provide formalized personal information for each scholar, such as professional experience, affiliation and position. Other than that, it also did analysis and visualize the publication topics historically and the academic cooperation network.



5 Questions to answer

5.1 How does the instructions cooperative publication look like?

- Cross institution cooperations are often strong connection. Names on the same paper are considered academia cooperators. And the relationship of cooperator can be viewed as a kind of social network links. Quite different from the one amount people in the same institution, the cross-institution cooperator relationship usually means stronger connection, due to the cost from the distance. Therefore, the cooperation among various institutions are stronger connections.

- If the institution is viewed as a node, and the cooperation is viewed as link, then the cooperators can actually act as context of the node and provide definitive information about the node. If the users are interested in one institution, then they are supposed to concern which ones it has cooperation with.

5.2 a. Which institution did better in publication by cooperating with other ones?

- Connections are usually positive indicators about how the institution is open to communicate, how strong it is connected in the academic network and how many potential publication opportunities it may own. Therefore, it is meaningful to investigate which institution did better in publication by cooperation.

- The rank could involve in many restrictive conditions. Sometimes the user might be interested in the historical ranking to estimate the evolution of the institutions.

5.3 Dose the cooperation strength has correlation with factors such as the distance between the institutions and the topics they are working on?

- According to the assumption that the cost of cooperation will increase as the distance is larger, the strength of the cooperation will decrease as the distance is larger.

- Some topics can be harder to work on independently than other ones, such as projects that requires supercomputing resources and large data collection.

6 Data Processing

6.1 Dataset

There are two datasets we need in this project. The first one is DBLP dataset and the second one is geographical dataset of each institution listed in the DBLP database. The second dataset would include two parts, first part is latitude and longitude of those institution in the world map and the second part is the name, number of publications and color data of each institution.
The original DBLP data is from https://dblp.uni-trier.de/xml/. However, the raw dataset is very large and difficult to clean. Fortunately, we found a cleaned dataset which contained all information we need from https://github.com/emeryberger/CSrankings. It is a summary of information for each article, which combines information of all authors.

6.2 Data Processing

With some transfomations, we got two datasets from the cleaned data. "collaborations.json" is an overview of collaborations between institutions with each element representing an institution and its collaborating relationships with ohter institutions. And "collaborationsDetails.json" contains more detailed information of conferences.

7 Visualization Design

7.1 Overview

There are three main parts in our visualization:

1. The overview visualization of collaboration relationships for all universities, which will be realized by a world map, a connection graph, and some line charts.

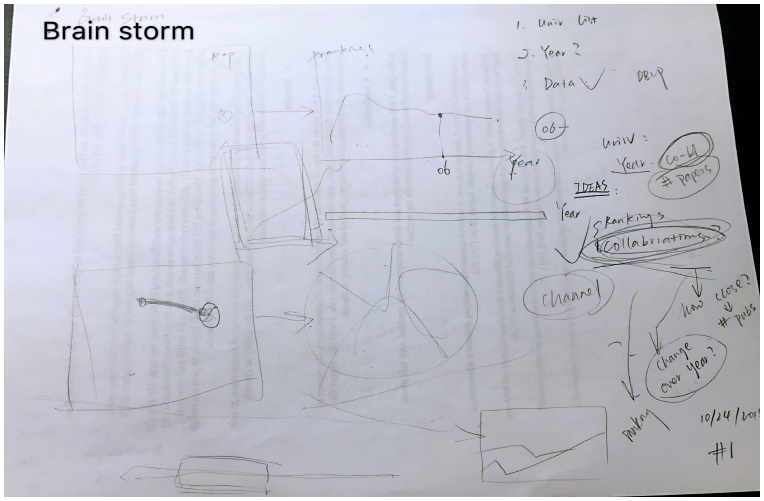
2. Visualization of basic information for one chosen university, which will be realized by an information box, containing the name, the CS-ranking, the number of collaborators etc. of this university, also a pie chart and a line chart indicating the related information in specific fields (AI, Sys, AL, and Interdisciplinary areas).

3. Visualization of comparison information among chosen universities, which will be realized by a comparison table.

7.2 Sketches

We started our design with brainstorming ideas.

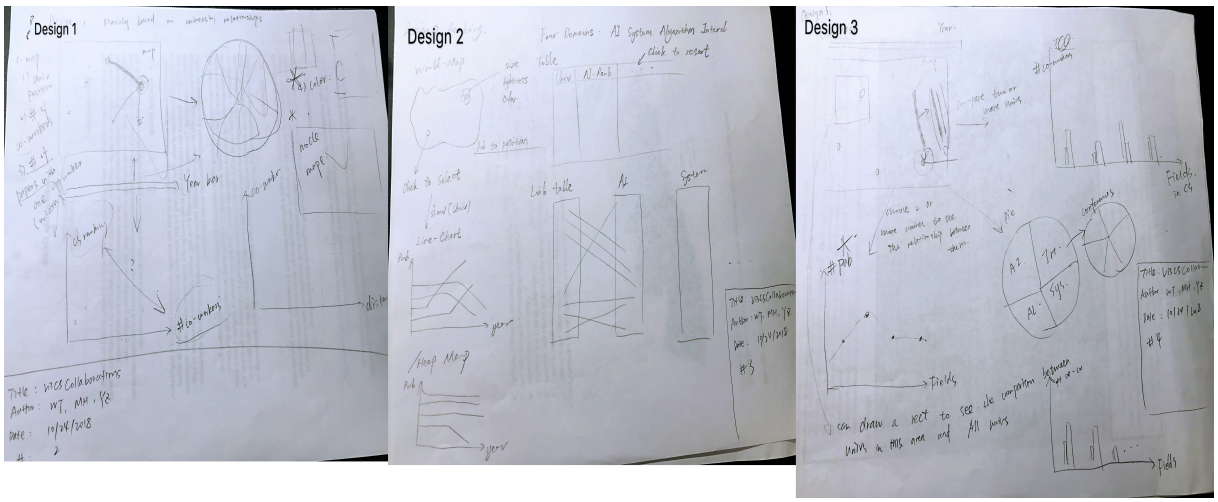
Then, we came up with three different design sketches.
In our first design, we mainly focused on how to display the overview of collaborations between universities with the change of years. We came up with a real map with circles, and a connected



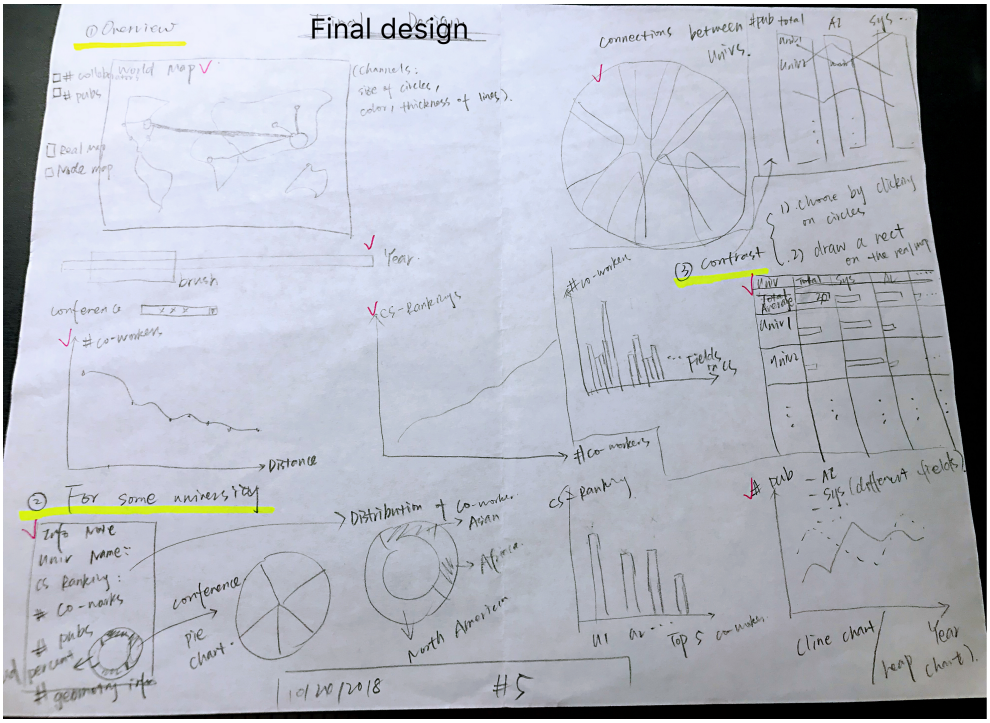
graph to show the connections between universities. Also, we tried to show the relationship between the number of collaborators and other attributes by some line charts. We considered to add an interactive 3D scatter plot too.

In our second design, we mainly focused on displaying the collaboration relationships in some specific areas in computer science, including AI, systems, theory and interdisciplinary areas.

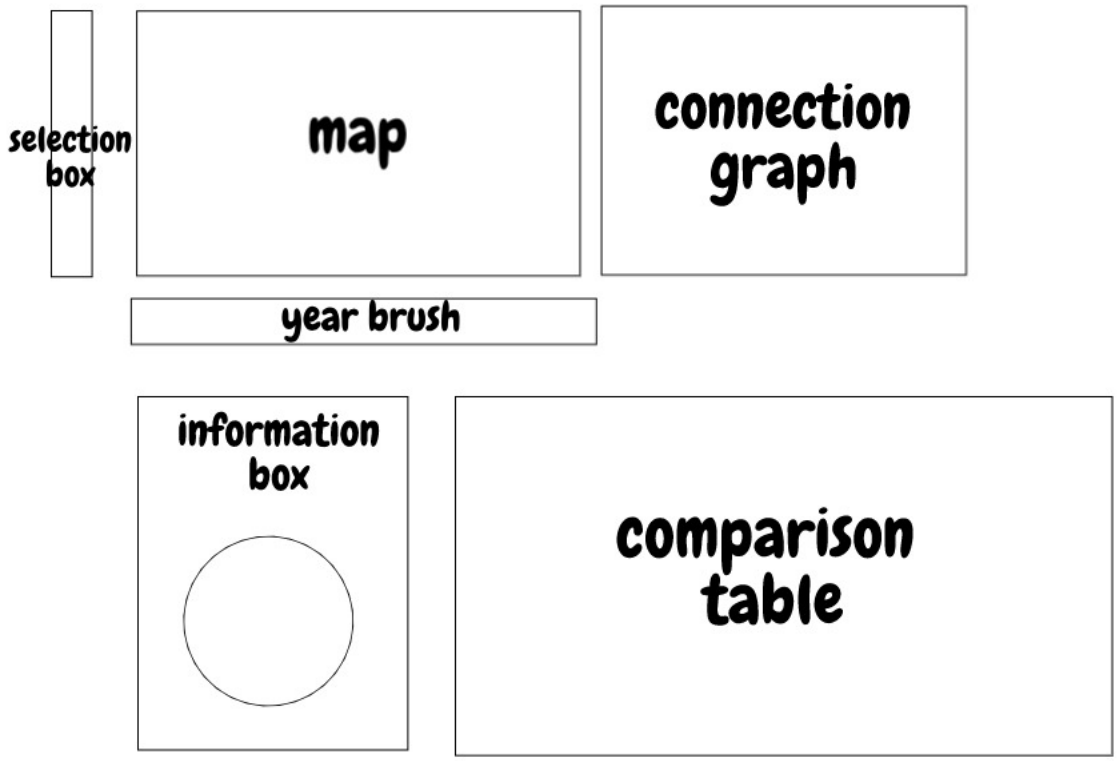
In our third design, we kept the world map, but mainly focused on representing the individual information and comparison information. Especially, we designed two ways of selecting circles. One is to click on the circle directly, and another is to draw a rectangle on the world map then circles in this rectangle will be chosen.



In the final design, we combined the advantages of these three designs and came up with a design including the three main parts stated above.



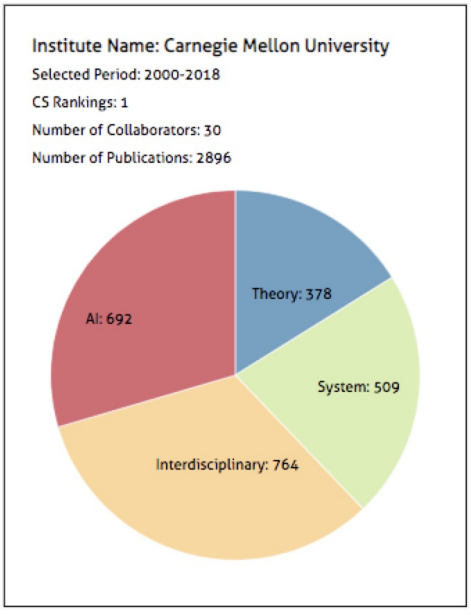
7.3 Prototype



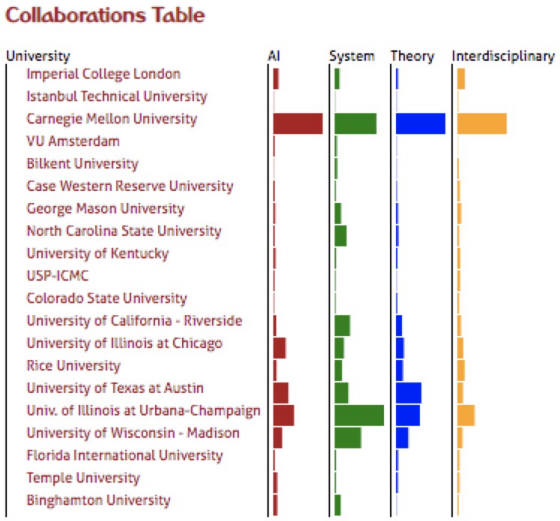
7.4 Design Evolution

1. **The map:** After implementing the world map and placing institutions on the map as circles, we found that circles are so dense in the United States and European area that it is hard to select a specifc circle. Therefore, we are considering to modify this view to display instituions only in the United States and their collaboration relationships with institutions inside and outside the United States.

2. **Information box:** This is an basic overview of a certain university, including university name, CS-Rankings, number ofcollaborators, number of publications. A pie chart to display the composition of its publications is also included in this box.



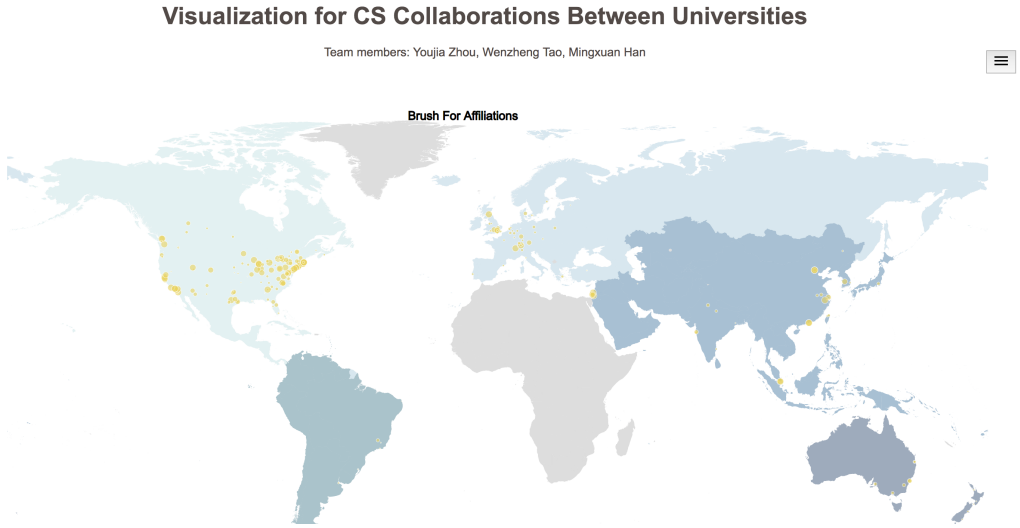
3. **Comparison table:** This is similar to the contrast table we built in hw5, which will show a bar char of the number of collaborators in each cell. Clicking the headers of this table will sort the whole table.



8 Implementation

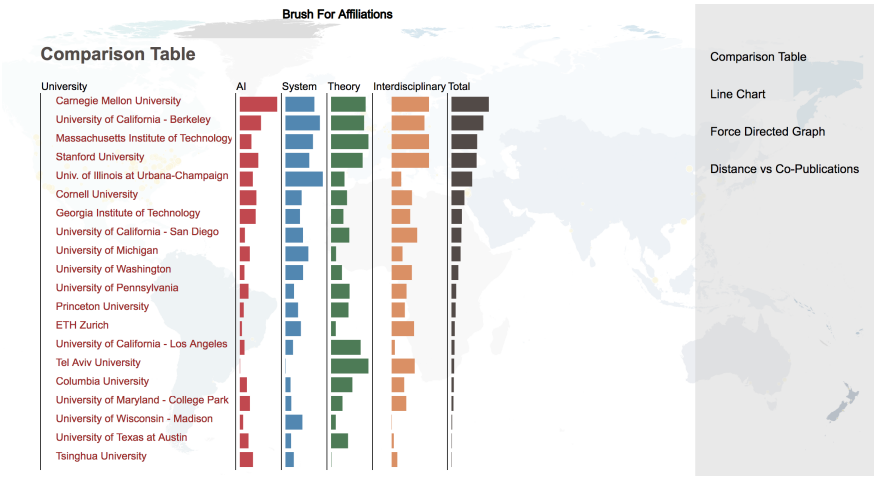
8.1 Overview

The worldmap with circles denoting the locations of the CS institutions.

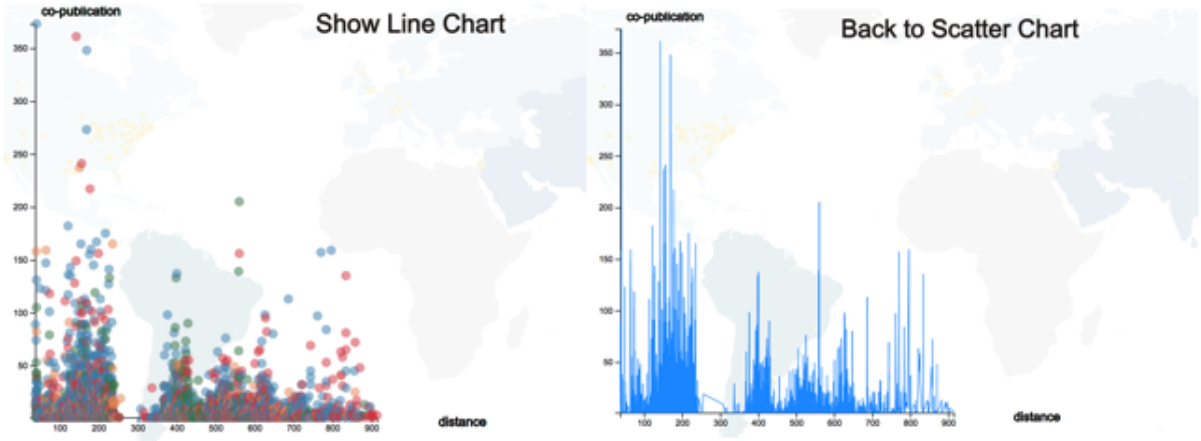


8.2 Contrast View

The institution table ranked by coolaboration publication amount.



Line chart of distance and co-publication

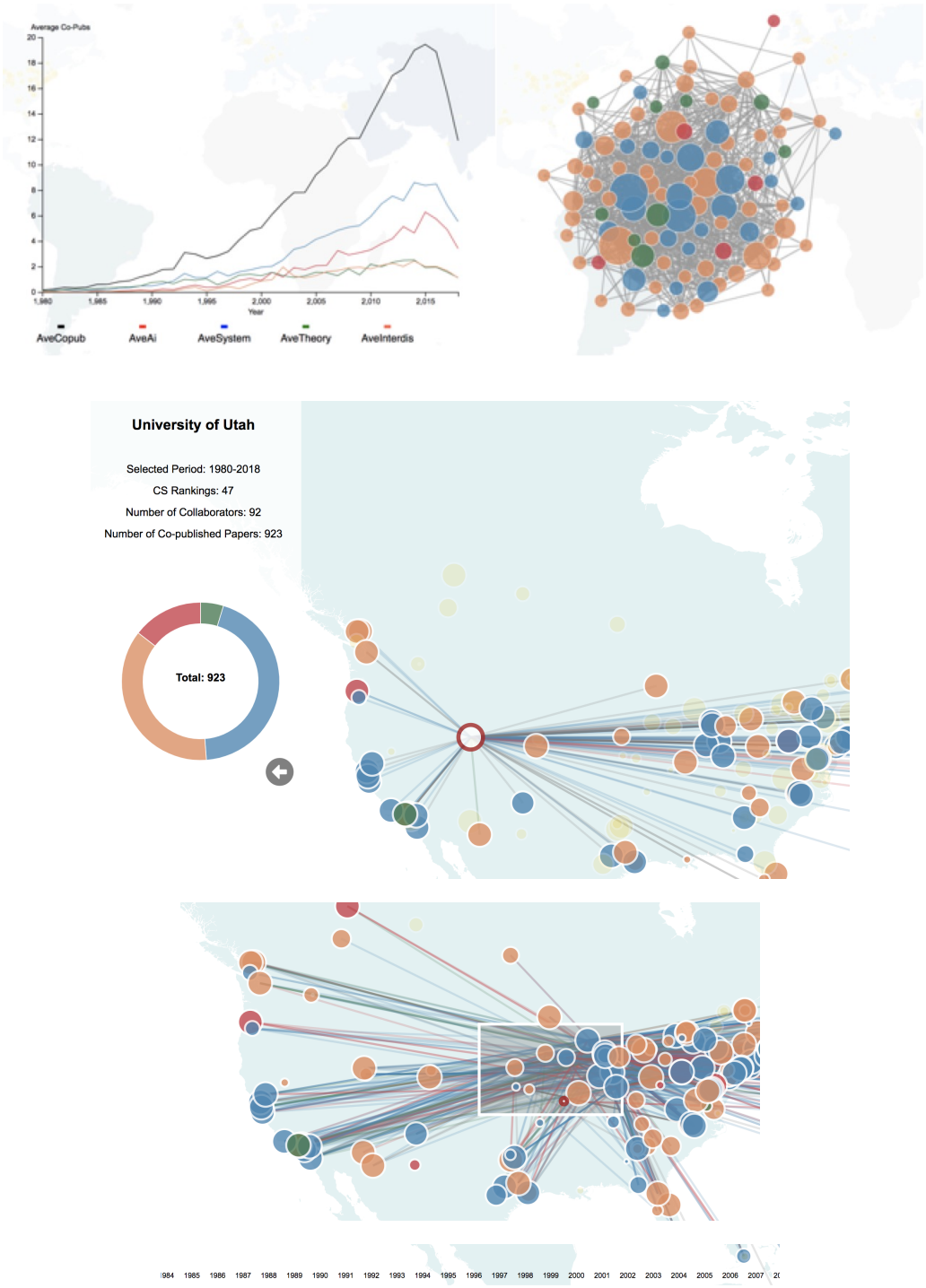


8.3 Certain university view

Displays the information such as university name, CS ranking, fields of research, number of cooperative universities and geographical parameters in a note panel. By click the certain field, the user will be able to see the ratio of the corresponding conferences in a ring chart.

8.4 Brushes

- Brush for Universities
- Brush for years



9 Evaluation

We preprocessed the data of cooperation among CS institutions. We could discover the phenomenon of clustering of cooperation with respect to distance and area such as AI. Therefore, the answer

is that there is correlation between distance, research topic and institution cooperation, while the influence of distance is more noticeable. The project could clearly display the cooperation network of the institutions. And the users are provided adequate freedom to filter the data to be displayed, by years and institutions, in various forms of scatter plot, line chart, ranking table and etc. It could be further improved with adding more features to the plots to display more concert data such as information detailed to the level of conference.