CS5346 Assignment 3 – CIRViz

Student Name	Bastian Morath
Matriculation Number	A0195628N

Introduction

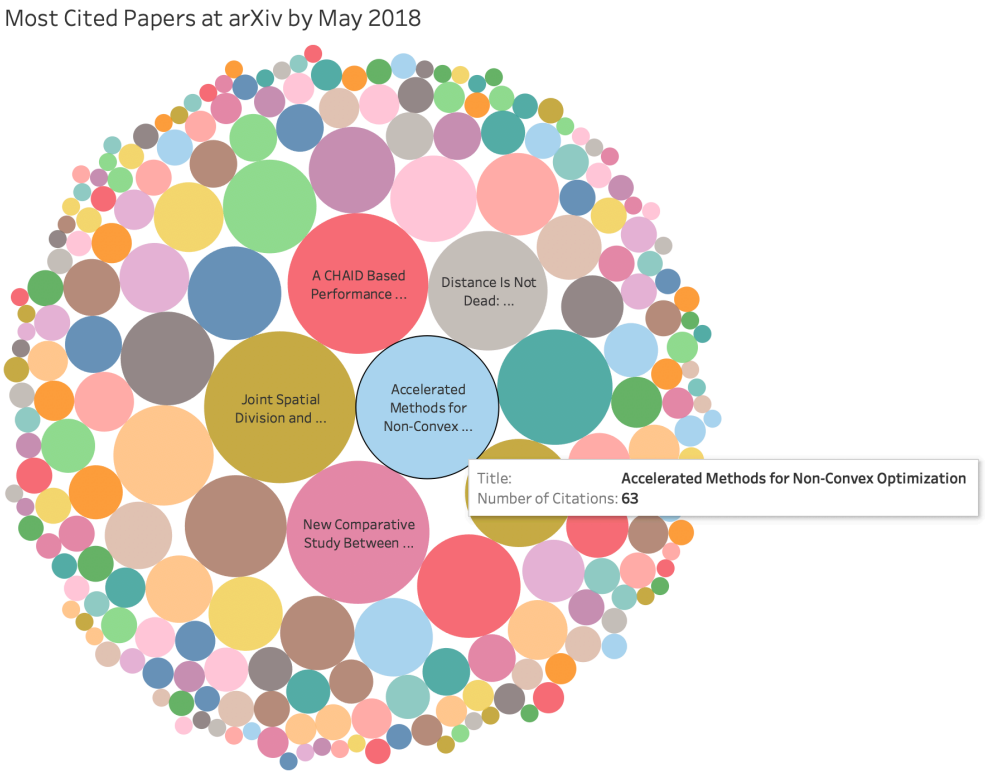
The aim of this project was to create interesting visualizations with *Tableau* about a large dataset containing research papers. The data comes from *Semantic Scholar*, a free, nonprofit, academic search engine. It contains a large amount of research papers (as of May 2018), together with attributes such as the paper title, the author-and co-authors and the number of citations. For our work, we extracted 200'000 papers from the full 39 Million available.
As discussed with Prof. Bimlesh, I worked alone and only did a subset of the tasks.

Visualization - Purpose & Method

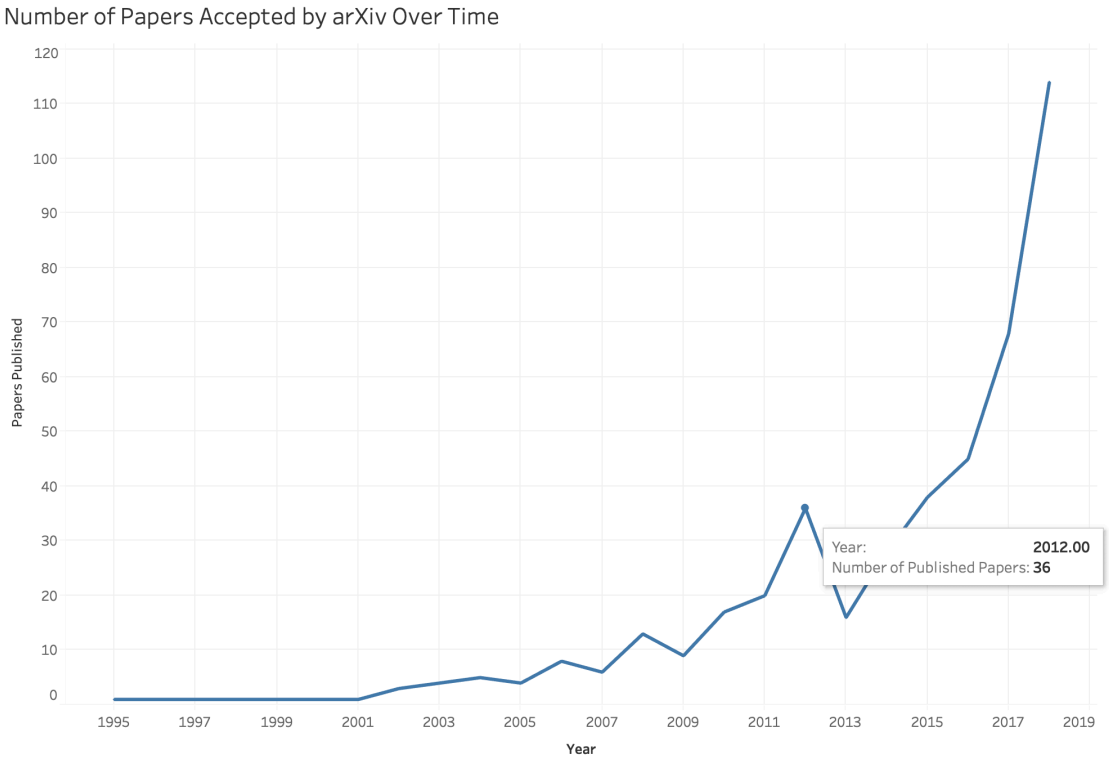
i)

Task	Visualization
2	Packed Bubbles
3	Graph
4b)	Scatter Plot

ii) Task 2: Top 5 most cited papers for venue arXiv.

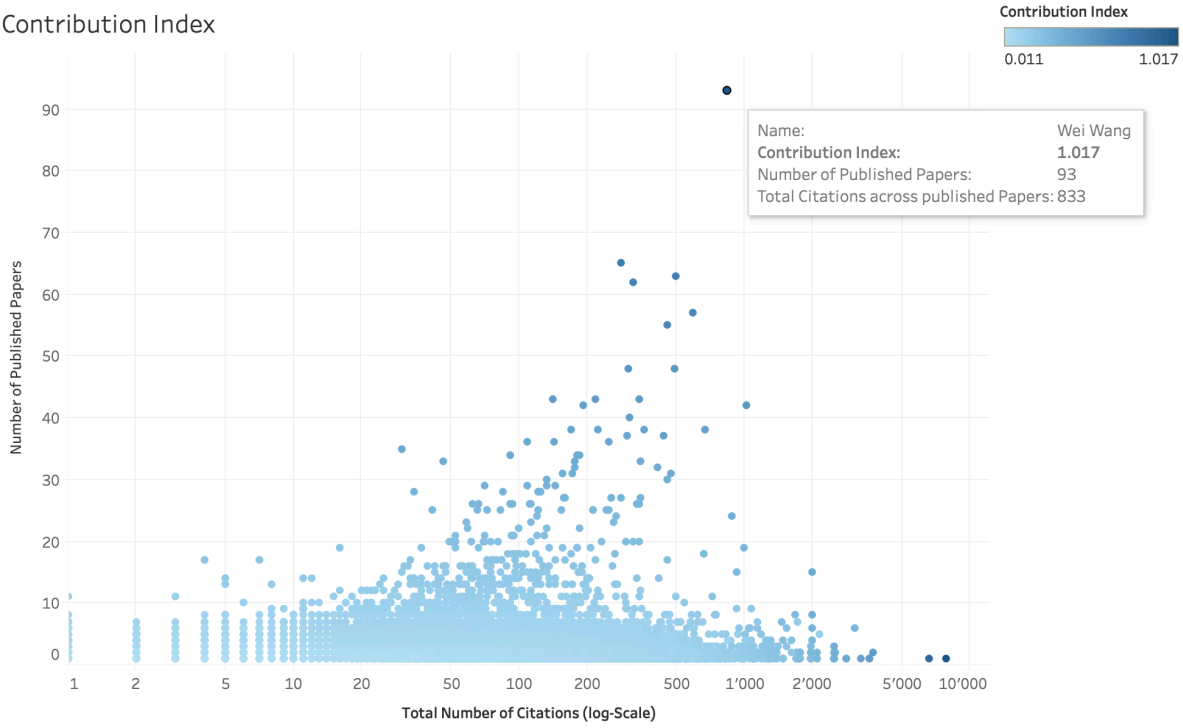


Task 3: Trend of publications across all available years for venue arXiv *



This visualization shows the number of accepted papers by arXiv over the years.

Task 4b) Contribution Index



The x-axis of this visualization shows the **total number of citations** the author has across all his published papers. Note the logarithmic scale.

The y-axis shows the **number of published papers** of an author.

The color of each bubble corresponds to the **Contribution Index,** a measure that is calculated as an average from both the number of published papers and the total number of citations.

* I decided to use arXiv instead of ICSE (as in the assignment), since my dataset had much more data about arXiv

iii) In detail:

Pre-Processing for all tasks:

When loading the JSON file into Tableau, one can choose which schema level to use, which determines the dimensions and measures available for analysis in the worksheets. I chose the top level plus the author name and the inCitation. I then hid all the fields that I did not use (such as the DOI and the URL). I also extracted the data locally. All those steps ensure that Tableau can handle the large amount of data better.

Note that Tableau automatically generates Indices for the different levels. For example, if a paper has 4 authors (and no citations), then Tableau creates 4 rows, each having a different author index from 1 to 4.

Task 2: Top 5 cited papers across for venue arXiv

I decided to visualize more than only the top 5 cited papers, but allowing to easily read out the top 5 papers from my visualization. I decided to use *Stacked Bubbles*, where each bubble corresponds to one paper that was published by arXiv. The number of citations (i.e. how many times other papers cited this paper) determines the size of the bubble. Therefore, bigger bubbles mean more citations. For the tooltip, I am showing the paper title and also the number of citations. I also included the paper title of the top 5 papers directly in the corresponding bubble.

Approach:

1. Create a bar chart
2. Filter the data to only show papers with venue *arXiv*
3. Use the *title*-dimension and *InCitations Index* dimension as row and column data, respectively. Change the InCitations Index to Measure->Maximum (I call this *num_papers* from now on)
4. Sort the titles along *num_papers,* and create a new set called *Top_5* from the top 5 papers
5. Change the chart to “Packed Bubble”
6. To only show label on top 5 bubbles: Create calculated bubble for the label with the following calculation: IF [Top_5] THEN [Title] ELSE “ “ END
7. Drag the title-dimension to the color-field
8. Adjust the tooltip to what you want to show

Task 4b): Contribution Index

I decided to base the contribution index of a user on two measures: The number of published papers, and the sum of all citations of the papers the author has published. For this I created a scatter plot, where the x-axis depicts the total number of citations, and the y-axis the number of published papers. I decided to use a logarithmic x-axis, since there were a few outliers that had a large amount of citations, but most did not. This would make it better to read the visualization in my opinion.

To emphasize the contribution index, I created a new calculated field *contribution_index*, which was then used to encode the color of each dot.

Approach:

1. Create a calculated field

   contribution_index = MAX([inCitations Index]) / 7767 + COUNTD([Document Index]) / 93,

   where the numbers are the maximum number of published papers and the total citations of an authors paper, respectively. I got the number from other charts, which one can quickly calculate and visualize in a separate helper sheet.

2. Create a calculated field

   total_citations = { FIXED [Name]: Sum({FIXED [Document Index] : MAX([inCitations Index])})}

   I used fixed functions, which basically allow to fix one dimension and then calculate a metric over it. I had to do this since for each paper, a new data entry was created for each citation (with a different inCitation Index). So for each distinct Document Index (i.e. paper), I calculated the Maximum inCitation Index, which corresponds to the number of Citations. I then sum this up over all papers of a particular author.

3. Create a scatter plot, using the *total_citations*-dimension and *Document Index* dimension as row and column data, respectively. Change the Document Index to Measure->COUNTD (I call this *num_papers* from now on)

4. Minor adjustments: Filter out names and total_citations that are null, change x-axis to logarithmic and adjust tooltip and axis names

5. Use the contribution_index as the color field