CS5346 Assignment 3 - CIRViz

March 31, 2019

Lu Wei	A0040955E
Julian Teh	A0163126M

Contents

1 Introduction

In this assignment we explore techniques involved in visualizing big data, including cleaning, filtering and transformation. We then use the processed data for exploration and extraction of insights. We split the work into two main portions to be divided between the pair: data exploration and coding of the final visualizations. For exploration we used Tableau with a small dataset to examine any interesting trends, and for the final visualization we used C3.js and D3.js to render with the full dataset.

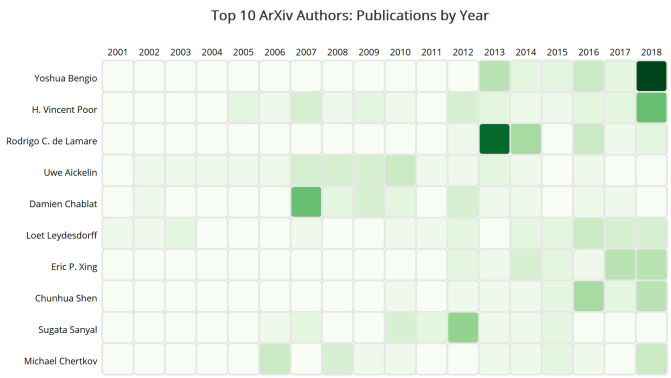
2 Visualizations - Purpose and Method

2.1

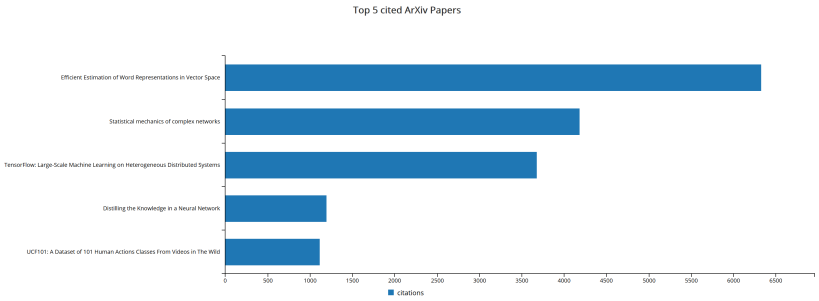
Task	Visualization
1	Heatmap
2	Categorical Bar Chart
3	Time Series Line Chart, Stacked Area Chart
4a	Force Directed Graph
4b	Tree Map

2.2

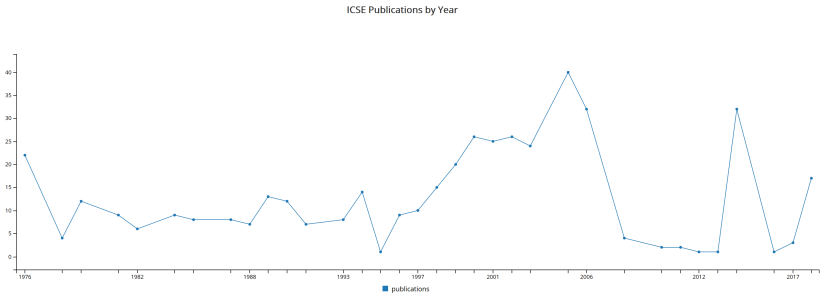
Preview images of visualizations



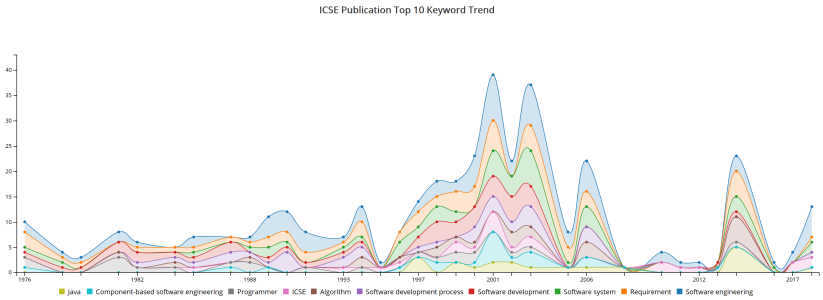
Task 1: Heat Map



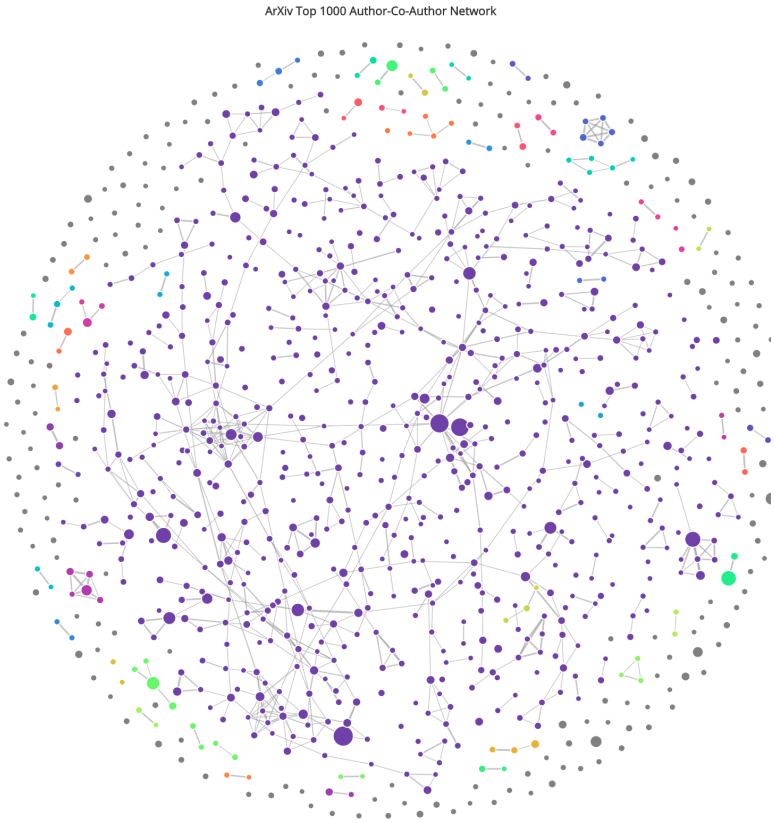
Task 2: Bar Chart



Task 3: Time Series Line Chart



Task 3: Stacked Area Chart



Task 4a: Force Directed Graph

Task 4b: Tree Map

2.3

Task 3b: ICSE Publication Top 10 Keyword Trend

Visual Encoding & Data Processing: We encoded the 10 most frequently occuring keywords of all time in ICSE publications into a stacked area time series chart. We first counted and extracted the 10 most frequently occuring keywords of all time, and then collected their frequencies by year. Using the X-Axis for year, and the Y-Axis for the number of occurrences of the keywords, we mapped out the points, and then connected them using a spline with colored area.

Interactions: Hovering over the points in the graph shows a tooltip including the year and exact counts of frequency of every popular keywords. Hovering over the legend highlights the area of the given keyword. Clicking on the keyword toggles to display and hide the corresponding visualization. Drag to zoom in and out of any specific time period of interest.

Task 4a: ArXiv Top 1000 Author-Co-Author Network

Visual Encoding: We represented authors as nodes in a graph, and drew the relations between authors and co-authors by connecting them with lines if they had written at least one paper together. The size of every node corresponds to the number of publications of the author. We defined the attractive force between authors to be positively correlated with the number of publications they have co-authored together; visually the thicker the link, the more two authors have collaborated. This gave us a force-directed graph representing the relation of authors. Note that each group of authors that are connected through coauthorship is colored the same and they are positioned close together in a cluster.

Interactions: Hovering over a node shows the author name and his/her total number of publications; hovering over a link shows the number of collaborations between two authors. Every node can be dragged out of place for better inspection.

Data Processing: There are over 47,581 authors in the dataset of 21,557 ArXiv publications we extracted from the first 10,000,000 entries of the full dataset. Plotting all of the authors in a network graph is not visually meaningful, therefore we sorted the authors by their number of publications and picked the top 1000 authors to visualize. The paper count per author is attached to every author object for visualizing node sizes. We also tallied the paper count per author pair for link visualization in the network. We labeled node groups using depth first search and then projected the groups into a rainbow color scheme. We also singled out the nodes without any connection (authors without co-authors) and colored them in grey to reduce visual noise.

Insights: Author Rodrigo C. de Lamare is very productive with 29 publications, but he only has one collaboration with one of the top 1000 authors, therefore he's a solo powerhouse. We can also see that there is one large group in the center (colored in purple) that connects majority of the top authors[1].

[1]Some nodes of the same color appear to be not connected to the rest of the nodes in the group because their co-authors that formed the links are not among the top 1000 authors

Task4b: ArXiV Author Contribution

Visual Encoding: We represented authors as cells in a treemap, and drew the relation between author and contribution using the size and color of the cell. Both size and color of each cell represent the relative number of publications made by the author, and the data is sorted in descending number of publications. This gives us a treemap that has cells of decreasing sizes from left to right.

Interactions: Hovering over a cell shows the author name and his/her total number of publications.

Data Processing: There are over 47,581 authors in the dataset of 21,557 ArXiv publications, and we use all of these to generate the treemap. We sorted the authors by number of publications, but did no other data manipulation to preserve the full scale of data so as to also preserve the insights we had observed.

Insights: Interestingly, we observed a phenomena that is well documented as Price's Law. Price's Law states that half the publications in any given field originate from a square root of all contributors. In our exploration of the data, we found that the relation between the number of authors and the number of papers was exponential in nature, and thus roughly half the publications are indeed written by a square root of the number of authors. In fact, when we observed this when we were exploring with only a subset of the data, but this held true when we used the full set of ArXiV authors, which illustrates that Price's Law holds even at scale.