Analysis and Visualization of Bigscholarly Data

September 15, 2018

1 Introduction

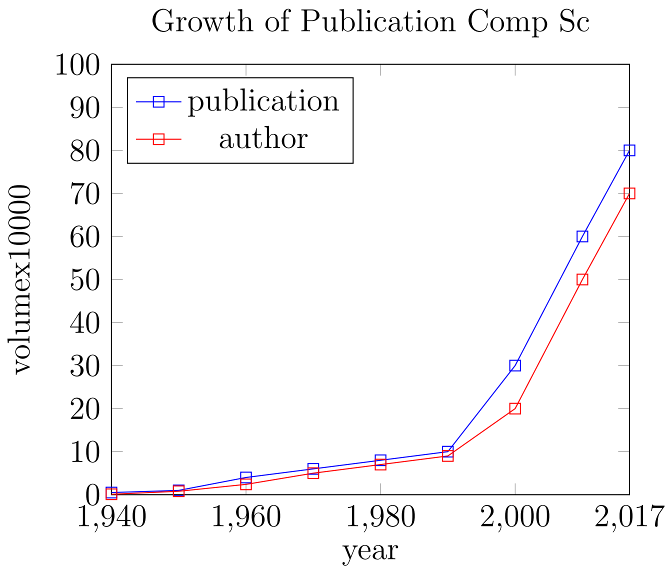
The objective of this project is to analyze scholarly publication data, comprising of close to millions of research publications, authors, citations and publication venues. The data is present in multiple tables, hosted in MySQL Cloud service. In this project we intend to build a web based application that presents various insights drawn from this dataset as visual network, charts, tables and heatmaps.

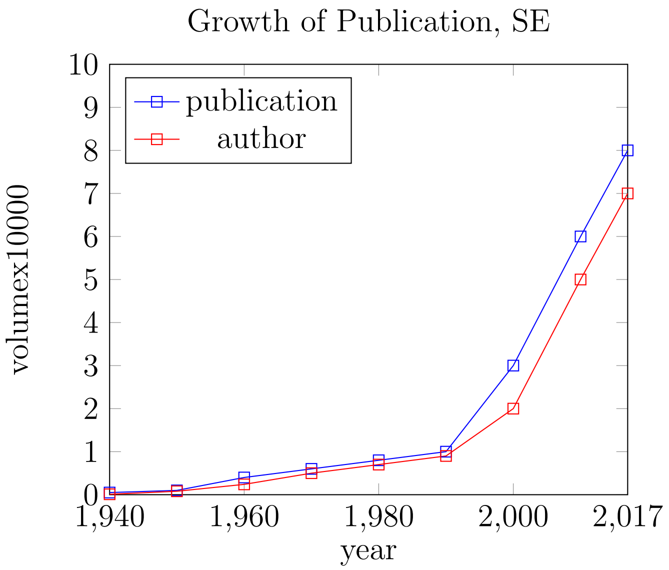
Charts, Graphs and Tables: The charts shown below are mere examples. For instance, do not assume that publication data starts from 1940 as shown in the graph. You have to query the database to fetch the earliest start date. Then use this start date in your subsequent queries.

Data Source There are two categories of tables. The table names that has "_AM" at the end are the big tables with papers from all domains of computer science. The table names that has "_SE" contains a subset of papers that belong to software engineering domain. The table names with "_AI" at the end contains papers in AI and machine learning.

2 Queries

1. **Publication Growth:** How many years did it take for the publication volume to double? The plot should look like this:


Growth of Publication Comp Sc

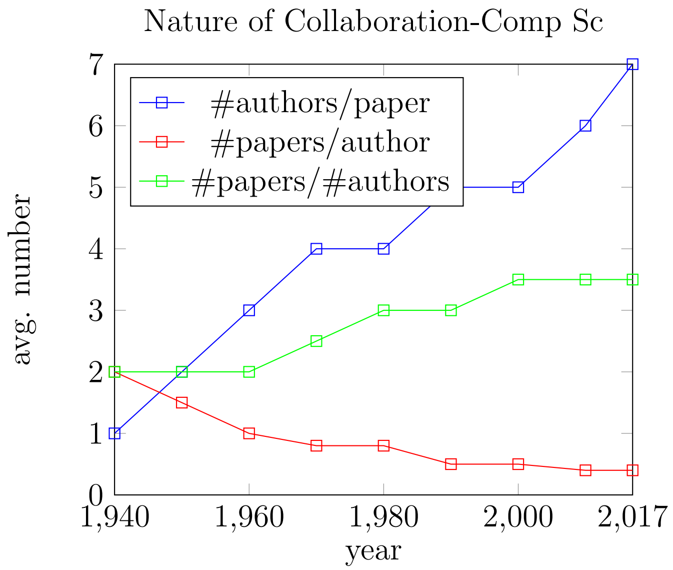

Growth of Publication, SE

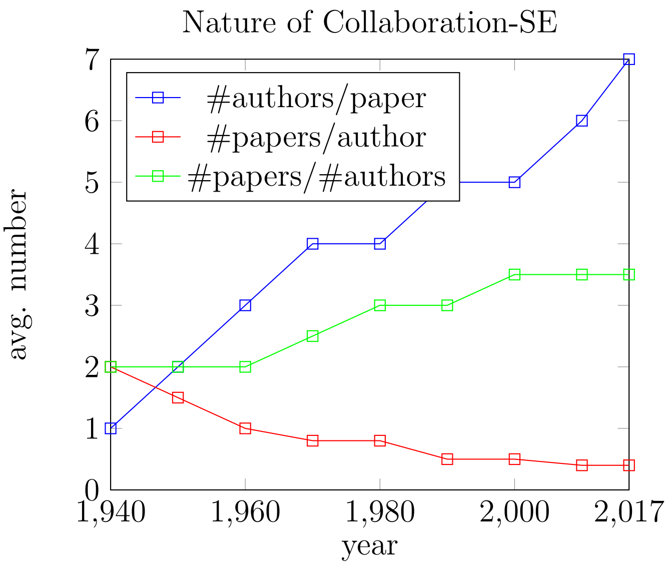
The example plot shows the number of publications (and the number of authors) in the Y axis. You need to plot this graph from the underlying dataset. Highlight the years when the volume is doubled. You should explore the power of the visualization package for highlighting. If nothing works, you can generate a table.

Furthermore, generate a table showing the top contributors (affiliations) to the volume of growth, in the following manner. Use the "_AM" tables only for this query.

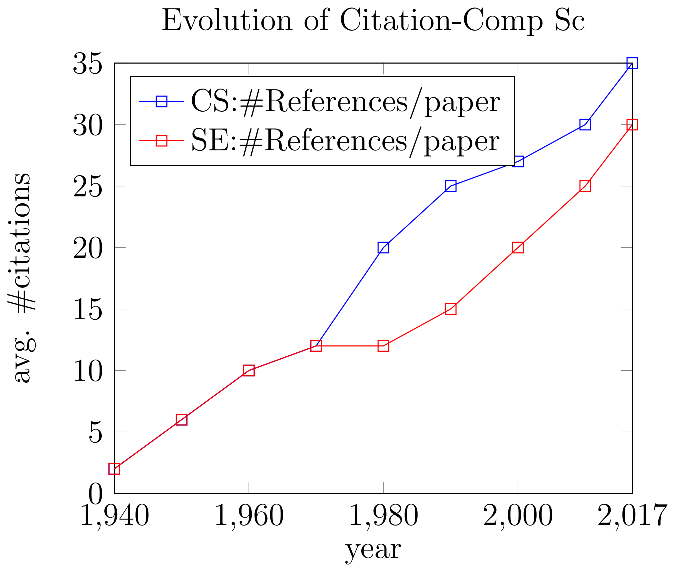
Affiliation	1940	1960	1980	2000	2017
Oxford Univ					
MIT Standfprd					

2. **Nature of collaboration:** In this assignment, you need to analyze the collaboration trend among the authors over the years. You need to plot three types of graphs- i) avg. number of authors per paper in each year ii) average number of papers published by an author each year and iii) total number of paper published in a year per author (#papers/#authors) The plot should look like this:

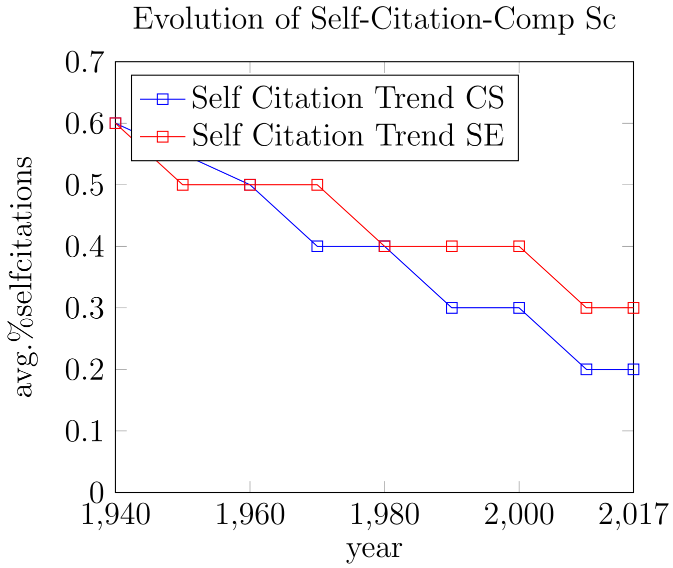

Nature of Collaboration-Comp Sc


Nature of Collaboration-SE

3. **Depth of Related Work Study:** How many references did an author cite in a published paper on an average? How is the trend over the years? For this, you need to compute the average number of citations per paper published in a given year and plot this average number across the years. Your plot should look like the following for Computer Science in general and SE in particular:

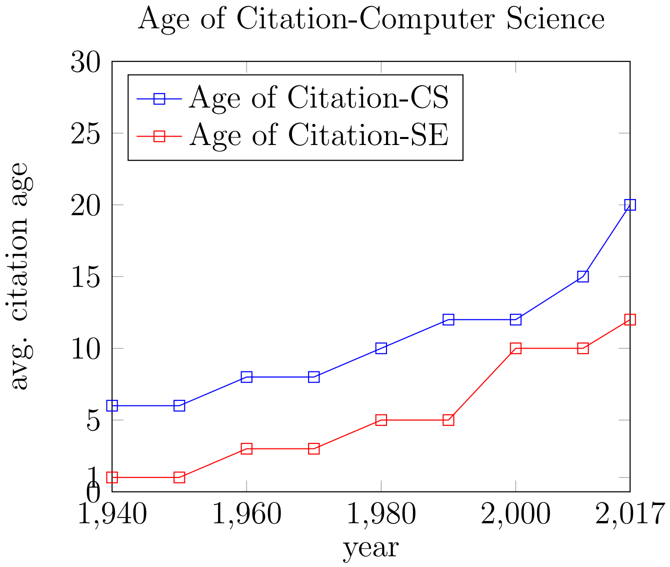


4. **Self-citation:** In this analysis, you check whether scientists are doing self-citations in an increasing manner. Typically self-citations are used to boost a person's own publication. Suppose that a paper p has cited a paper p1. This citation is a self citation if there is at least one author common between p and p1. Once you mark a citation to be a self-citation, the next step is to compute the %self-citation. We compute this as follows. Suppose that p has N citations, out of which M are self-citations. Then the %self-citation is $\frac{N}{M}$. You take the average of the %self-citations across papers published in a given year. Then you plot the result. The result should look like the following for computer science in general and software engineering in particular:



5. **Myopic vs. Deep referencing:** How many years back did authors look back in the published

literature? You can compute this by checking the age of a citation. For a paper p published in a year y, select the year of publication of a cited paper p1 in p (say y_c) and compute $(y - y_c)$. Then compute $\max_{y_c}(y - y_c)$. Do this for all the papers p published in a particular year y, and take the average. Plot the graph as follows for computer science in general and SE in particular:


Age of Citation-Computer Science

END