# Visualizing Collaboration with Superstars

## A Novel Approach to Visualizing Collaboration

**Preston Hull**[1]
**Supervisors: Hayley Hung, Vandana Agarwal, Chenxu Hao**
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Preston Hull
Final project course: CSE3000 Research Project
Thesis committee: Hayley Hung, Vandana Agarwal, Chenxu Hao, Megha Khosla

**Abstract**

Superstar researchers - those who author research papers which are far more widely cited than average - are generally well-respected within their fields and are frequently sought by new researchers for advice on career development and for collaborations. Though the effect of collaboration with superstar researchers on associated researchers is not well-researched yet, it is believed that collaboration with superstar researchers increases the research output of associated researchers, but may decrease their originality and innovation. This project explores a method of visualizing the career development of researchers associated with superstars. Using a dataset of authors, papers, and paper citations, a graph has been created with papers as edges and nodes as authors to visualize the career development of associate authors after collaboration with a superstar. The utility of this visualization is evaluated using heuristics.

# 1 Introduction

The network of institutions and researchers conducting and publishing papers is vast and well-established. The research on collaboration between authors, particularly involving those who write the 0.01% most cited papers, however, is not. Research output of those associated with these so-called superstar researchers, when adjusted for quality, has been found to by 5-8% after the sudden death of the superstar [2]. Another potential negative effect of the inequality and hierarchy in academic research is that funding accumulates towards the researchers who are already most successful – an effect known as the Matthew Effect. Researchers who are successful during their early research grants, receive more funding later in their career and apply for grants more often than those who are not[4]. Primarily, this project is inspired by the article by Kelty, et al., *Don't follow the leader: Independent thinkers create scientific innovation*, which explores the potential side effects of collaboration with superstars in academia, and finds that, overall, academia "pays a price by focusing resources and attention on superstars" [10]. However, there is a lack of effective visualization of the effect of superstars on associated researchers. With all of the potential issues caused by the hierarchical structure of academia, it is important to research this further. Therefore, the primary research question this paper seeks to answer is how the career growth of researchers who collaborate with superstars can be visualized in terms of their research, independent of that of the superstar.

This task introduces some important subquestions to include:

- Which dataset is most appropriate for this project?
- How can a visualization from an appropriate dataset be produced?
- What are some useful conclusions that can be derived from the visualization and the dataset?
- How can the usefulness of the visualization be properly evaluated?

Graph visualization is a common problem both in industry and in academics, and as such, there are many available open-source solutions for graph libraries. It was decided to create an interactive graph in a web browser (as opposed to static images) to allow for manipulations, zooming, and panning on the graph. Many JavaScript graphing libraries are based on D3.js[1], but for this project, it does not provide enough abstraction and its direct use would be more difficult than necessary. Cytoscape[6] is a very popular, mature, and well-featured graph visualization and graph theory library used by many

[1] `https://d3js.org/`

universities and major companies. Based on the available alternatives, I chose to use Cytoscape for this project.

There has also been significant research into graph visualization techniques and usability. For this project, it is essential to ensure that the visualization produced is useful and clear. Bennett (et al.)[3] describe a number of important usability heuristics for graphs, including limiting edge crossings, limiting edge bends, and reducing edge lengths (while still trying to create uniform edge lengths). Additionally, Herman (et al).[8] explain that graph visualization, in the form of nodes and edges, is most useful if there is an inherent link between the data. Therefore, it is important to ensure that the authors used as nodes in the produced visualization are well-connected and that the graph design is intuitive.

This paper is structured as follows: the second section describes the problem in greater detail and the methods used to answer these research questions; what follows is a section describing the derivation of key parameters used in the creation of the graph; the next section explains the results of my research and shows the key visualization; then there is a section about responsible research and ethical implications of my research and its results; and finally, a section concluding my research and deliverables.

# 2 Problem Description and Collaborative Work

To visualize career development of those associated with superstars, it is important to first define the desired output of the visualization. This section describes the process of conceptualizing the graph, identifying an appropriate dataset, and collaboratively pre-processing it in preparation for building the graph visualization.

## 2.1 Formal Problem Description

This project aims to explore potential methods to visualize the career development of those who associate with superstar researchers, independent of the superstar themselves. Further, it aims to produce and evaluate such a visualization on the basis of usability criteria. To that end, this project produces a graph of scientific research. Through discussions with my supervisor and responsible professor, we have collectively built an understanding that this involves graphing "superstar authors" (the most highly cited authors in a given field), "associated authors" (those who collaborated with superstar authors), and "collaborators" (those who later worked with an associated author). Furthermore, the graph must be readable and useful, must not contain too much extraneous information, and must be properly evaluated.

This graph requires several pieces of data. Since authors are nodes and papers are edges, at least a list of papers and paper authors from multiple disciplines are necessary. Additionally, a database of citations between papers is necessary to determine superstar status. Thus, we chose to find a dataset including authors, papers, and citations. My project group collaborators needed a similar dataset, so we decided to collaborate on the selection and processing of a suitable dataset.

Processing such a dataset also has specific challenges. This dataset is large, so it is necessary to efficiently store and process at least a few hundred gigabytes of data. It will also be necessary to efficiently relate records, for example, to relate citations and papers to find the most highly cited papers, and relate those papers to authors to find the most highly cited authors.

## 2.2 Collaborative Work with Research Project Team

My project group and I began with the selection of a suitable dataset. Our search began with the Semantic Scholar Open Research Corpus (S2ORC) originally used by the paper by Kelty (et al.)[11]. This dataset is delivered in a large set of files conforming to the JSON Lines[2] format in 4GB partitions, totaling to around 1 terabyte of data. Each record of the dataset contains the full text of a paper and lists of indices in the text of where various attributes like the authors, citations, author affiliations, and the abstract. To produce useful results, these indices must be processed into their original text and then related with that of other papers. To that end, I wrote a simple script that took the S2ORC dataset as an input and produced a semi-processed JSON Lines output containing, for each paper, a list of authors and the other attributes (using the indices from the original dataset on the full text). After some refinement and tuning of regular expressions, I was able to produce a fairly consistent result.

Once the S2ORC dataset was processed, our project group considered how to process this data. In the case of a processed JSON Lines dataset, relating papers to other papers through their citations would require, for each paper, linearly scanning through the entire dataset to find each paper that it cites, and incrementing a counter on that paper for each citation. This has a time complexity of $O(k * n^2)$ for all papers, where k is the number of papers each paper would cite on average, and n is the number of papers, which is impractical for a dataset of this size. We also considered using language models to determine the discipline of the paper and to parse metadata, but we arrived at the conclusion that even an execution time of one second, which is very generous, processed over millions of records, leads to a processing time of at least 12 days, and likely at very high cost as well due to the cost of API calls. Our team then began to look for other, more efficient options.

We looked to how datasets of this size are processed in industry and found a few potential solutions. One of them is to use Apache Hadoop[3] and Apache Spark[4], a software framework and engine respectively used in industry for big data analytics. Typically these applications are run on a cluster and take advantage of parallel computing. In our case, we wanted to ensure that our data processing was simple to replicate, and as such we chose to avoid this option to avoid the expense of running a cluster computer. We also considered using ElasticSearch and Kibana, a big data search engine and companion UI respectively, for processing this dataset, but they do not offer the join functionality we needed, and the time required to familiarize ourselves with the interface and API was too significant. In the end, we chose to run a PostgreSQL[5] database for processing our dataset, because it is a mature project, has all the necessary functionality for examining the dataset, and because it is heavily optimized for joins and operations on a single-machine cluster. Every optimization our team considered, such as binary data storage, indexes, and sorting, are already implemented in PostgreSQL, and in a way that is likely far more efficient than our team could accomplish during the scope of our research project.

After further examination, our group found that the S2ORC dataset is not appropriate for our use case because we do not need the full text from the papers and have had trouble normalizing various aspects of the dataset. For example, author names cannot easily be related because of different spellings or abbreviations of their name (think of J. Doe vs John Doe). We found that Semantic Scholar offers a dataset more suitable for our use case called the Semantic Scholar Academic Graph (S2AG)[6]. The S2AG contains

[2] `https://jsonlines.org/`
[3] https://hadoop.apache.org/
[4] https://spark.apache.org/
[5] https://www.postgresql.org/
[6] `https://blog.allenai.org/semantic-scholar-academic-graph-for-developers-6188cfec84d4`

| | Table Name name | Total Size text | Data Size text | Index Size text | Other Size text | Record Count real |
|---|---|---|---|---|---|---|
| 1 | citation_context | 472 GB | 432 GB | 35 GB | 4362 MB | 1.6440422e+09 |
| 2 | citation_entry | 211 GB | 141 GB | 70 GB | 39 MB | 2.5091238e+09 |
| 3 | paper_entry_paperfield | 139 GB | 112 GB | 27 GB | 32 MB | 3.585438e+08 |
| 4 | paper_entry | 96 GB | 90 GB | 5839 MB | 25 MB | 2.0964885e+08 |
| 5 | paper_entry_paperauthor | 60 GB | 39 GB | 21 GB | 11 MB | 6.0305574e+08 |
| 6 | author_entry | 9082 MB | 6499 MB | 2582 MB | 1832 kB | 8.876748e+07 |

Figure 1: PostgreSQL database table sizes

several datasets, including abstracts, authors, citations, papers, and TLDRs (summaries of papers generated by language models). Additionally, this dataset is already neatly organized and indexed with unique identifier for each paper, citation, and author. Loading this database into a PostgreSQL database makes it fairly easy to join on these identifiers, without having to normalize other attributes such as paper titles or author names. Our team decided to use this dataset.

The datasets including authors, citations, and papers contain hundreds of gigabytes of data. We began loading the authors dataset first since it is simple and does not contain multi-variate attributes that require multiple tables. This dataset was also used to develop a script to process s2ag datasets into the database. Though the authors dataset contains no multi-variate attributes, the papers dataset contains multi-variate attributes for both authors and fields (paper disciplines). Additionally, for citations, it was decided to place citation contexts (the paragraphs where the citations appear) in a separate table from citation entries because they are not needed in many cases and excluding them may improve query performance. Therefore, we have six tables: $citation\_context$ (which contains the paragraphs where citations are located), $citation\_entry$ (which contains a record of all citations between papers), $paper\_entry\_paperfield$ (the fields and disciplines of each paper), $paper\_entry$ (the main paper database, listing all papers), $paper\_entry\_paperauthor$ (a list of all authors for each paper), and $author\_entry$ (which contains various attributes of each author). A script was written to organize the s2ag datasets in this manner, and then optimized through, among other things, batching insertions (multiple insertions at once instead of the naive approach of just one) and multi-threading. The database processing took several days. The resulting sizes of our database tables are shown in Figure 1.

# 3 Determining key parameters for data visualization

To produce a useful and intuitive visualization of such a large dataset, it is important to ensure that the right subset of the dataset is used. To that end, I have decided on several key parameters to be used to filter the dataset. A key parameter I have decided is to focus specifically on recent data, to ensure that the visualization is as relevant to the present as possible. The data used for the visualization will focus on the years 2003 - 2023 (the last 20 full years available at the time of writing), which should ensure that the conclusions presented by the visualization are recent. In terms of determining superstar status, the authors of the papers which received the greatest number of citations between 2003 and 2013 were used, and from the top of the list, three were selected per discipline, with six disciplines total. The three selected were offset by five from each other (as in, positions 1, 6, and 11) to ensure a fair sampling from the list and to avoid situations where the three superstars were collaborators on

the same highly-cited paper. Additionally, to further clarify the graph by removing unnecessary information, only a sample of associates and collaborators are made into graph nodes. The formula chosen for this purpose is, where $N$ = the number of associates or collaborators, $ceil(0.1N) + 5$, which ensures that an author with more associates or collaborators is still shown with more connected nodes, while ensuring that the authors with many hundreds of connected nodes do not crowd out the rest of the graph. To identify *associated authors* (those who are associated with a superstar), collaborations between the previously-identified superstars and other authors between 2008 and 2013 were considered. To identify *collaborative authors*, collaborations between the associated authors and others between 2013 and 2023 were considered. This produces a large list of associated and collaborative authors, with authors as nodes and papers as edges.

The quantity of six disciplines was chosen to provide a reasonably broad overview of the development of associated authors in various fields while still providing a reasonably small graph. The disciplines themselves were chosen to be a broad selection of STEM and non-STEM subjects, and also to vary in relative size and prominence. The chosen disciplines include Art, Computer Science, Engineering, History, Law, and Physics.

To properly signify the superstars, associates, and collaborators, the superstars are shown as red nodes, the associates as green nodes, and the collaborators in grey. The sizes of the nodes depends on the prominence of the author (the number of citations they received - as defined earlier), relative to the prominence of the field (measured as the average citation count of the top 10 cited authors between 2003-2013, using the same dataset as before). Edges which represent more papers (more frequent collaboration between two authors) are thicker than edges which represent only one paper. Within a discipline, nodes are clustered and appear in a colored region of the graph corresponding to that discipline.

# 4 Data Processing, Visualization, and Evaluation

This section describes the process of producing graph data from the intermediate dataset, creating a visualization, deriving conclusions from the data, and evaluating the visualization.

## 4.1 Data Processing

The process of converting the dataset from authors, papers, and citations to a graph dataset of nodes and edges was complex and arduous due to the size of the dataset. Processing the dataset brought several questions, including what the dataset result schema should be, how to efficiently query the PostgreSQL database, and how to write the script that efficiently queries the database and produces the resulting dataset. Each one of these steps caused significant challenges.

To produce a usable graph from the dataset, it is necessary to first consider the desired output. The graph shall contain a visualization of authors as nodes and papers as (undirected) edges. Each edge could contain multiple papers. Nodes should be colored according to the category of author (superstar, associate, or collaborator). The size of the output should not be excessive, both in file size and in number of edges and nodes, and, ideally, the output can be converted into the input format for multiple graph libraries.

Then, I considered a potential resulting schema for the graph data. Initially, I reviewed academic research describing graph data structures used for large graph databases. Particularly, I considered the semi-structured tree model described by Angles and Gutierrez[1]. However, I came to the conclusion that any graph data structure too large to fit in a simple JSON file is also too large to properly visualize. Thus, I decided to save the

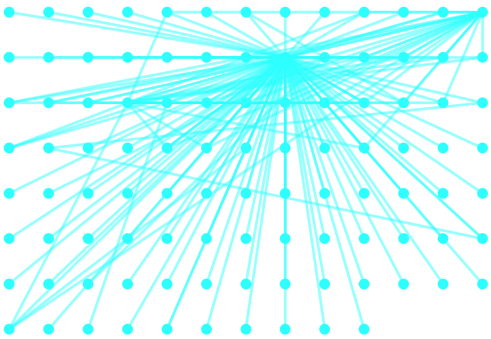

Figure 2: Cytoscape's grid layout

graph in a JSON file. I also considered how to properly address the problem of undirected edges potentially being stored twice. To solve this, edges are stored in a nested JSON object, with the lower-value author ID being on the outside and the higher-value author ID being on the inside. This means that there is only one way to store an edge. Each edge contains a list of papers that the two authors collaborated on. Additionally, when this file is loaded into JavaScript or Python, a hash map object can be used to efficiently query edges. Nodes are stored as an object as well, indexed by author ID and containing a few pieces of information about that author, such as their citation count, their name, their discipline, and their type (superstar, associate, or collaborator). This JSON data structure is simple, efficient, and appropriate for this use case. However, for usability purposes, it has been decided to split the graph into six different graphs, one for each discipline. Therefore, the output should be a JSON file for each discipline.

With a schema and a plan, I then began writing a script to convert the dataset into the JSON data structure. The first attempt was the naive method of querying the database for all associates and then, for each one, all collaborators. This results in several thousand queries, each taking an average of around 30 seconds, which would result in multiple days of execution time and several gigabytes of storage. However, optimizations previously used in the database insertion script, such as multi-threading, cannot easily be used in this case because nodes connect to other nodes and it is difficult to ensure thread safety. The solution was to move the database querying, which is the bottleneck, to worker processes while performing graph operations on one process. Another optimization used is to cache intermediate results from database queries. Through these optimizations, the script execution time is around three hours, which is enough for this purpose.

## 4.2 Results: Creation of the visual graph

Once the dataset was processed, I began to write a small web application using the Cytoscape library to host the visualization. This web application should be lightweight, minimalistic, and prominently present the visualization. To avoid the use of unnecessary libraries, this visualization is written in pure HTML, CSS, and JavaScript. I added a tab bar, much like those present in web browsers, to allow the user to easily select which discipline to view. To minimize memory usage, each graph is loaded when the user clicks on the button to view it and unloaded when the user clicks on another button. These choices result in a clean, fluid interface.

An important consideration in the creation of the graph was the layout of the nodes. By default, Cytoscape uses a grid layout, which places nodes in a rectangular grid, as shown in Figure 2. This layout obscures the edges and violates several of the heuristics

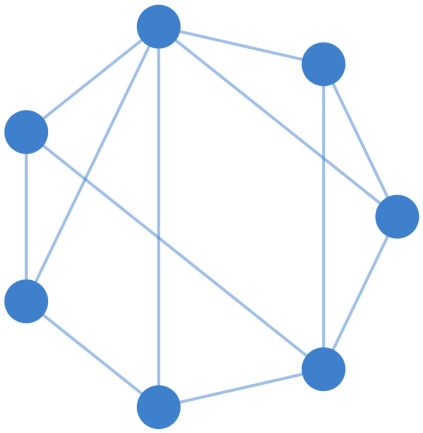

Figure 3: Cytoscape's AVSDF layout

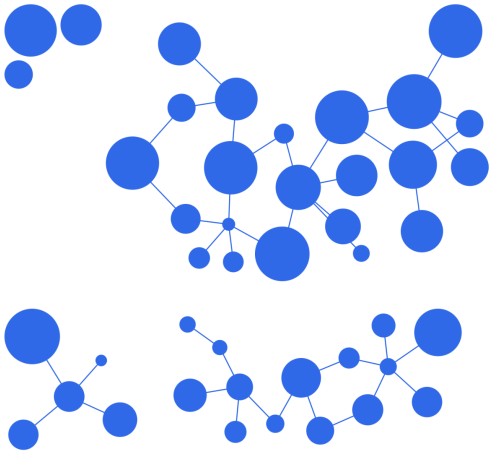

Figure 4: Cytoscape's cose layout

defined by Bennett (et al.)[3], particularly related to edge placement. Another layout supported by Cytoscape is AVSDF, based on a circular drawing algorithm authored by He and Sykora[7]. This algorithm, shown in Figure 3, aims to minimize edge crossings. However, when run on a dataset of the size visualized in this project, it creates a large ring with edges crossing the inside of the ring. This does not properly show the collaborations between associate authors and collaborators. In addition, Cytoscape supports force-directed graph layouts, such as cose (Compound Spring Embedder). cose implements the algorithm defined by Dogrusoz (et al)[5]. It was determined through trial and error that this layout produces qualitatively the best graph layout, particularly when accounting for graph heuristics. Therefore, it was chosen to use the cose layout to produce the graph.

The graph itself varies widely between disciplines. The art discipline has only one associate and zero collaborators, while the physics discipline has relatively many associates and collaborators. Table 1 describes the number of associates and collaborators considered per discipline, along with contextualizing information about their relative

| Discipline | Prominence | Associates | Collaborators | Collaborators per associate |
|---|---|---|---|---|
| Art | 1 | 1 | 1 | 1 |
| Computer Science | 20.1 | 77 | 10356 | 134.5 |
| Engineering | 10.7 | 55 | 4281 | 77.8 |
| History | 1.6 | 24 | 3993 | 166.4 |
| Law | 2.1 | 56 | 32109 | 573.4 |
| Physics | 33.6 | 177 | 109698 | 619.8 |

Table 1: Associates, collaborators, and collaborators per associate, per discipline, derived from the first, sixth, and eleventh most popular author per discipline (as described previously)

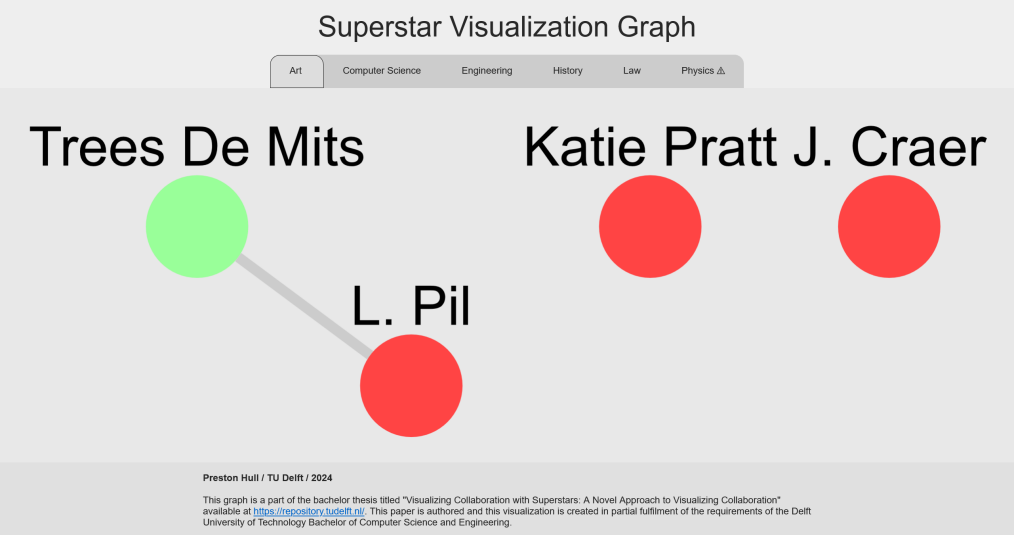


Figure 5: The superstar visualization graph for the Art discipline

prominence[7]. Figure 5 shows the graph for art, and figure 6 shows the graph for computer science. The graph visualization tool has been published to GitHub Pages[8], a static web-hosting service commonly used to host sites for open-source projects.

## 4.3 Analysis of results

The graphs and data show notable results, particularly when comparing disciplines. When looking at the graphs, it is clear that associates in some disciplines, particularly law and physics, generally have more collaborators after working with superstars than associates in others. For example, in the art discipline, among the sampled data, there was only one associate and one collaborator (the superstar themselves), leading to just one collaborator per associate, while the physics discipline has 109,698 collaborators and 177 associates, leading to an average of 619.8 collaborators per associate. This result for physics means that authors are generally very successful after working with superstars, whereas for art, this effect may be less significant. There does not appear to be a strong correlation between the relative prominence of a field and the number of collaborators per associate.

[7]The number of citations of the most cited author in the field, given previous constraints, divided by the number of citations of the least prominent analyzed field.

[8]`https://technophilus.github.io/rpgraph/`

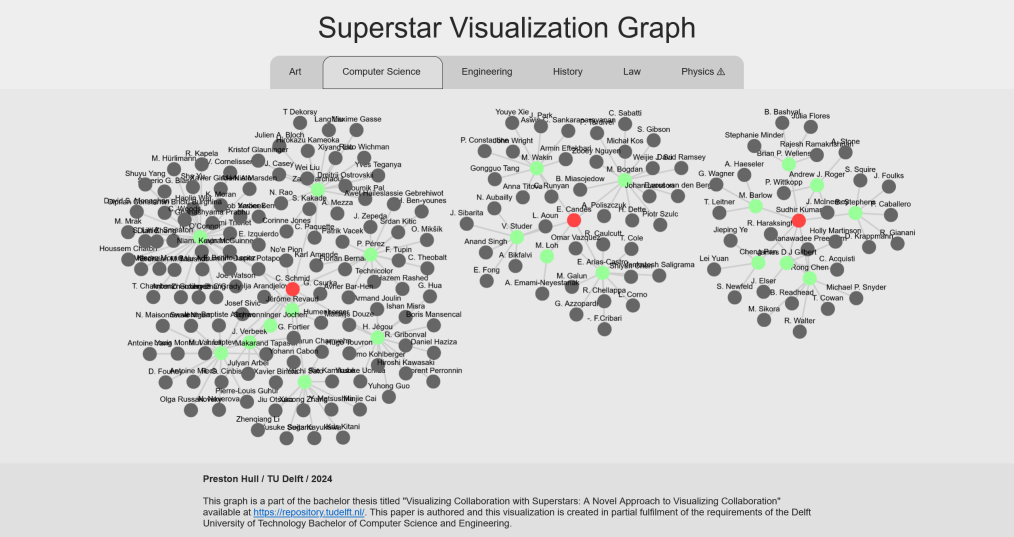


Figure 6: The superstar visualization graph for the Computer Science discipline

### 4.3.1 Evaluating the visualization against heuristics

To produce an effective and understandable visualization, it is important that it is intuitive and understandable. Bennett (et al.) describe heuristics for nodes, edges, and the general layout of the graph that can be used to measure the usability of a visualization. In this section, I will apply these heuristics to the superstar visualization graph to evaluate the usability of the graph.

One important heuristic is that nodes should be evenly distributed throughout the graph and there should be some distance between the nodes. Bennett (et al.) describe that when clustering nodes, "the distance between nodes in a cluster should be equal, and the number of different distance levels between nodes should be minimized." The superstar visualization graph fulfils this criterion for smaller graphs, such as Art and Computer Science, but for larger graphs, such as Law and Physics, it struggles with the quantity of nodes and places them closer together. Another notable node heuristic is that nodes should be kept away from unrelated edges. This graph adheres to that heuristic fairly well for smaller graphs but again struggles for larger graphs.

Edges and edge placement also have relevant heuristics. According to Bennett (et al.), "By far the most agreed-upon edge placement heuristic is to minimize the number of edge crossings in a graph. The importance of avoiding edge crossings has also been extensively validated in terms of user preference and performance. Similarly, based on perceptual principles, it is beneficial to minimize the number of edge bends within a graph[3]". The force-directed graph layout algorithm used to create the superstar visualization graph minimizes edge crossings and does not create bent edges, even for very large graphs. Additionally, edge length should be minimized to reduce the size of the graph. This superstar visualization graph does this appropriately when accounting for the previously mentioned node heuristics regarding space around nodes. Overall, the graph visualization tool adheres to these edge heuristics for all size graphs.

Finally, it is important to consider heuristics relating to the overall layout of the graph. An important heuristic is ensuring that the shape of the graph is similar to the shape of its container. In this case, that would be the shape of the frame of the viewer web page. The graph is dynamically generated and, as such, should appropriately be generated with a shape similar to its container. There are a few other heuristics, such

as maximizing convex faces, which are not achievable in this graph due to its nature as a tree-like graph (cycles exist but are rare). Therefore, this graph visualization tool generally adheres to layout heuristics.

# 5 Responsible Research

As with any research project, ethics and the ability to replicate findings are important for proper scientific research. To that end, I have evaluated potential risks involved in the handling of the dataset used for this visualization, the processing of the dataset, the visualization itself, and the potential use cases of the visualization. Additionally, I have considered potential difficulties in replicating this process. In this section, I will detail my findings.

## 5.1 Ethical considerations

A critical concern that was considered during the production of this visualization is respecting the privacy of the authors, the licenses of the papers, and the license of the original dataset. Semantic Scholar, the organization responsible for the S2ORC and S2AG datasets, requires an API key for accessing the dataset, though they provide API keys with barely any preconditions. Therefore, I do not consider the dataset to be private or restricted. Semantic Scholar also clarifies in their license agreement[9] that the data provided from the dataset typically are granted under separate licenses, including the CC BY-NC license and the ODC BY license. These licenses require attribution and the CC BY-NC license prohibits commercial use[10]. The papers provided as part of the dataset are typically under other licenses. This project does not use the contents of the papers, only the S2AG datasets, and therefore the CC BY-NC and ODC BY licenses are applicable. The use of these datasets for this project is compliant with these licenses.

Another important consideration is the set of conclusions that may be made from this visualization. It is important for those using this visualization to draw the correct conclusions from it, and not to perpetuate existing unfair biases. For example, it is known that race and gender, for example, are known to affect academic success, as shown by Johnson-Bailey and Cervero[9]. There is a significant risk that those using this visualization could make unfair connections between academic success and race, and then attribute such success as a characteristic of the race of the author. However, attempting to mitigate this by, for example, normalizing the size of the nodes by the race of the author presents other significant ethical risks, such as determining how much it is fair to adjust the node size and whether it is reasonable to increase the size of the nodes of authors who are of racial minorities simply because of their race. Therefore, it is left up to the interpreter of the visualization to understand that there may be racial or other biases in the presented data and to account for that when drawing conclusions from the visualization.

Environmental effects of large computation tasks are an additional factor that have become even more noticeable with recent advancements in commercial applications of machine learning, such as ChatGPT and Google Gemini. It has been noted that cloud computing consumes significant amounts of energy that have detrimental environmental effects. According to Uchechukwu (et al.), as of 2014, datacenter power consumption worldwide is estimated to account total 26 gigawatts, which was 1.4% of global energy consumption[12]. To ensure a limited impact and reasonable execution time based on available resources, the scripts running the visualization were designed to run as efficiently as possible, and the choice of using PostgreSQL for our dataset storage also

[9] `https://www.semanticscholar.org/product/api/license`

[10] `https://creativecommons.org/licenses/by-nc/4.0/deed.en`

reduces energy consumption by ensuring that queries and accesses are as efficient as possible. However, it is worth considering that replications of this processing could be done inefficiently and result in more power consumption than necessary.

### 5.2 Reproducibility

It is critical that the results of scientific research can be reproduced by others. Unfortunately, the size of the dataset naturally causes difficulties in replication. Those who wish to replicate this processing would require, at minimum, one terabyte of storage and a persistent server (for PostgreSQL). For reasonable performance, an SSD is also required, preferably one with a high IOPS (I/O operations per second) specification. It is common for tasks like this to be performed on rented cloud servers. At the time of writing, using the AWS pricing calculator[11], assuming a 1 terabyte SSD storage volume and an ec2 (virtual server) instance with 4 vCPUs and 16 GiB of memory, the cost of processing this dataset for two weeks would be USD$81. This could be unaffordable, particularly for student researchers. This could pose a significant hurdle to reproducing the data processing and visualization.

The various scripts used for data processing and storage have been published at the 4TU[12] data repository[13] and all use only the python standard library and psycopg2[14], a mature library used for interfacing with PostgreSQL. This is an intentional design choice intended to improve reliability and ensure that the creation of this visualization can be replicated. Therefore, the risk is low that libraries are deprecated or no longer maintained. However, there remains a risk of the deprecation of standard library functions or compatibility-breaking changes in future Python versions.

It is possible that processing the dataset could also be hampered by changes in the s2ag dataset schema, or by the elimination of the dataset entirely. Namely, it is possible that Semantic Scholar is no longer hosted by the Allen Institute for AI[15]. Use of a different dataset may be difficult due to changes in schema or fundamental changes in the structure of the data. Therefore, the risk of being unable to replicate this data processing and creation of the visualization due to changes in the dataset or lack of availability of the original dataset is fairly high.

## 6 Conclusions and Future Work

This report demonstrates a procedure through which the careers of those who are associated with superstar researchers can be visualized and how the resulting visualization can be evaluated through established guidelines and heuristics. However, the visualization and the analysis of the results of the visualization can both be further developed. This report presents only one possible way to produce an evaluate such a visualization, others are described in the following subsection. Overall, there is a significant amount of development that can be made in this direction.

### 6.1 Future work

There are several directions for future development of this project, particularly relating to data set selection, efficiency and scale of data processing, visualization technologies and techniques, and analysis of results. More specifically to the process described in this

[11] `https://calculator.aws`
[12] https://www.4tu.nl/
[13] `https://data.4tu.nl` with DOI 10.4121/4243dace-9bc2-4ca2-a343-3d481d6a9316
[14] `https://pypi.org/project/psycopg2/`
[15] `https://allenai.org/`

paper, future development could focus specifically on more intelligent sampling from the dataset (rather than simply the first, sixth, and eleventh elements) and closer adherence to the graph heuristics.

S2AG by Semantic Scholar was the original dataset used for this project and was evaluated to contain all of the information within the scope of this project. However, if additional specific data is needed or if more interdisciplinary knowledge is required, a different dataset may be more useful. An in-depth exploration into superstar collaboration may use something like S2ORC, a dataset of full texts and publication information of papers, to produce a more specific and accurate dataset.

This project used a PostgreSQL database as an intermediate step when processing the dataset. For the approach used in this project, and at the scale it was used during this project, this was a good choice. The PostgreSQL database requires a persistent server and a high-performance storage drive, and in our use, it was restricted by the I/O performance of this drive. When attempting to produce larger graphs, use a larger portion of the dataset, or when attempting to use a non-persistent server (such as DelftBlue or a comparable supercomputer) to process this dataset, there are alternatives, such as Apache Spark and Hadoop, that may be more suitable.

Web technologies, such as Cytoscape, JavaScript, HTML, and CSS were chosen to produce this visualization in an interactive format. However, Cytoscape has far more advanced configuration and performance tuning options than the ones used during this project. The graph could be, for example, made to be better animated or better able to align with the graph usability heuristics described in previous sections, such as by placing nodes further apart and making the labels more clear. For this project, a warning label was placed next to the physics tab to warn that it may take a significant amount of time to render. In the future, this could be resolved through performance improvements.

Various conclusions have been drawn from the visualization, but far more research can be conducted into specific results from the visualization and what factors may lead to them. For example, research can be done to understand why physics has, on average, far more collaborators per associate within this sampling of the dataset than art. Additionally, more intelligent sampling can be done to ensure the observed results more closely match reality. The reasons for these patterns are likely to be domain-specific and require extensive research into each discipline to fully understand.

## 6.2 Conclusion

The visualization of such a broad subject such as the career development of academics, complicated by a very large dataset, is not a trivial problem. It is especially not trivial to condense such a dataset into a visualization that is usable and intuitive. Producing this visualization takes many steps that each require consideration, such as finding an appropriate dataset, evaluating methods to process this dataset, processing the dataset into a form from which it can be queried, processing the dataset into a graph dataset, actually creating the graph, and ensuring it is presented in a form that is usable and understandable.

From this visualization, it is possible to draw notable conclusions about the career development of those who work with superstars, particularly across disciplines. In some disciplines, such as physics, those who associate with a superstar author visually appear to have more success than those in other disciplines, such as art. There is not a direct correlation between the number of collaborators per associate and the prominence of the field, which may mean that some fields are naturally more collaborative than other fields.

Overall, this process of visualizing academic career development generally works well, creates a useful graph, and can lead to several interesting conclusions, but it falls short

in terms of fair sampling, visualization performance, node placement heuristics, and conclusions about the academic field made from the data, and all of these issues can be addressed in the future as further development on this project.

# References


[1] Renzo Angles and Claudio Gutierrez. An introduction to graph data management. *Graph Data Management: Fundamental Issues and Recent Developments*, pages 1–32, 2018.

[2] Pierre Azoulay, Joshua S Graff Zivin, and Jialan Wang. Superstar extinction. *The Quarterly Journal of Economics*, 125(2):549–589, 2010.

[3] Chris Bennett, Jody Ryall, Leo Spalteholz, and Amy Gooch. The aesthetics of graph visualization. In *CAe*, pages 57–64, 2007.

[4] Thijs Bol, Mathijs De Vaan, and Arnout van de Rijt. The matthew effect in science funding. *Proceedings of the National Academy of Sciences*, 115(19):4887–4890, 2018.

[5] Ugur Dogrusoz, Erhan Giral, Ahmet Cetintas, Ali Civril, and Emek Demir. A layout algorithm for undirected compound graphs. *Inf. Sci.*, 179(7):980â994, mar 2009.

[6] Max Franz, Christian T. Lopes, Gerardo Huck, Yue Dong, Onur Sumer, and Gary D. Bader. Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics*, 32(2):309–311, 09 2015.

[7] Hongmei He and Ondrej Sỳkora. New circular drawing algorihtms. *Proc. ITAT*, 4:15–19, 2004.

[8] Ivan Herman, Guy Melançon, and M Scott Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on visualization and computer graphics*, 6(1):24–43, 2000.

[9] Juanita Johnson-Bailey and Ronald Cervero. Different worlds and divergent paths: Academic careers defined by race and gender. *Harvard Educational Review*, 78(2):311–332, 2008.

[10] Sean Kelty, Raiyan Abdul Baten, Adiba Mahbub Proma, Ehsan Hoque, Johan Bollen, and Gourab Ghoshal. Don't follow the leader: Independent thinkers create scientific innovation. *arXiv preprint arXiv:2301.02396*, 2023.

[11] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online, July 2020. Association for Computational Linguistics.

[12] Awada Uchechukwu, Keqiu Li, Yanming Shen, et al. Energy consumption in cloud computing data centers. *International Journal of Cloud Computing and Services Science (IJ-CLOSER)*, 3(3):31–48, 2014.