SCALABLE DATA VISUALIZATION METHODS FOR ACADEMIC CAREERS

A Dissertation

Presented to

the Faculty of the Department of Computer Science

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Kyeongan Kwon

July 2016

SCALABLE DATA VISUALIZATION METHODS FOR ACADEMIC CAREERS

Kyeongan Kwon

APPROVED:

Dr. Ioannis Pavlidis, Chairman
Department of Computer Science
University of Houston

Dr. Zhigang Deng
Department of Computer Science
University of Houston

Dr. Guoning Chen
Department of Computer Science
University of Houston

Dr. Brian Uzzi
Kellogg School of Management
Northwestern University

Dean, College of Natural Sciences and Mathematics

I would like to give my deepest gratitude towards Dr. Ioannis Pavlidis, who has been guiding me through my research. He has given me numerous opportunities throughout my PhD years, and always shared his views and experience in several of the projects that I worked upon. The amount of knowledge and experience I have gained from him is priceless.

I want to thank all my committee members, Dr. Zhigang Deng, Dr. Guoning Chen, and Dr. Brian Uzzi for their advice and feedback making this research possible. Also, I thank all lab members for being my colleagues and friends; especially, Dinesh Majeti, whom I have been working closely with.

Last but not least, my family is the backbone of my success. Being able to continue my education in America was only a dream. And to this very day, I cannot believe I was able to fulfill my dream. It is all thanks to my family. Although we are miles away and seas apart, my family made me feel as if I were home. I cannot be more thankful for their unconditional love, encouragement, motivation, and support.

SCALABLE DATA VISUALIZATION METHODS FOR ACADEMIC CAREERS

An Abstract of a Dissertation

Presented to

the Faculty of the Department of Computer Science

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Kyeongan Kwon

July 2016

Abstract

In this dissertation, I have developed scalable data visualization methods to depict a scholar's accomplishments at a glance. The evaluation of scholarly achievements in academia is largely based on the researcher's publication record. This record is communicated in exhaustive detail in the researcher's curriculum vitae (CV) or in summary via her/his h-index. The h-index, although a convenient abstraction, does not consider neither the time of the publication nor the impact factor (IF) of the journal where it appeared. I present a novel method that visually complements the h-index, revealing at a glance the nature of a researcher's scholastic record. This method (which includes the visualizations Scholar Plot and Academic Garden) is particularly appropriate for web interfaces, as it produces information that is compact and simple, yet highly illuminating.

Scholar Plot uses Google Scholar, Impact Factor and NSF/NIH/NASA funding data to create a temporal representation of a researcher's publication/funding record that blends publication prestige with paper popularity and funding information. Scholar Plot affords an insightful appraisal of academics at one's fingertips. Academic Garden applies to individual academics, departments, colleges, and any other academic group thereof, such as a research lab or a project team. Academic Garden uses the flower metaphor to visually articulate performance of academic entities. The width of the flower's stem is commensurate to the academic funding the entity received ('juice conduit'). The height of the flower's stem is commensurate to the impact of the entity's intellectual products ('visibility'). The diameter of the flower's disc is commensurate to the prestige of the venues where these products

appeared ('fancy factor'). Scholar Plot and Academic Garden bring clarity, transparency, and fairness in hiring, promotion, tenure, and funding decisions.

For the validation of the Academic Garden, I ran data analysis using Endowed Chaired Faculty, a prestigious honor in the United States, for the top 10 universities according to the US News Report 2015 [19]. The analysis demonstrated that chaired faculty can be predicted using the 3 merit criteria of citations, impact factor and funding.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A curriculum vitae (CV) provides a synopsis of an individual's achievements. The CV content varies by profession. Academic CVs feature prominently a publication section. This section references the researcher's journal papers and other scholarly products.

Search, promotion, and award committees that screen CVs go through lists of publications trying to form opinions about the candidates' records. Does candidate A or B have enough publications? Are they of high quality? Did they have any impact on the research community? In a highly competitive context, these questions do not always have clear answers. Another question that needs to be addressed is whether the candidate has been funded. If so, has the candidate done justice to the amount of funding obtained? This also enables one to decide if the candidate's output is in proportion with the input.

Chapter 2

Related Work

There have been some work on the quantification of academic careers, focused on a quest for a 'number' that sums up an academic's scholarship. The most well-known outcome of this line of research is the h-index, proposed by Hirsch [13]. A scholar has an index of h if s/he has published h papers each of which has been cited in other papers at least h times (Figure 2.1).

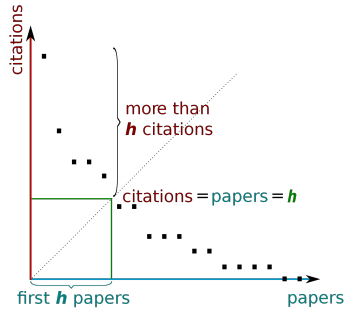

Figure 2.1: h-index from a plot of decreasing citations for numbered papers

The h-index depends on both the number of publications and the number of citations. Hirsch demonstrated that h can predict honors, such as the National Academy membership and the Nobel prize. He also suggested that it could predict advancement to tenure, although with some uncertainty. Despite its value, the h-index has weaknesses and when used, context should be carefully taken into account; such context includes the academic field and the academic age of the candidate [2].

With the advent of Google Scholar [11], information about a researcher's publication record and her/his h-index has become easily accessible. Then, with the ease of access of the internet, this information has become ubiquitous.

In this dissertation, I introduce data visualization methods that complement publication information contained in a standard CV and summarized by the h-index. The tool produces a temporal visualization that connects the h-index with the paper citations and the journal impact factors along with the funding data.

There have been other efforts in visualizing patterns of scientific production and impact [4, 5, 17]. Recently, a mobile app (DBIScholar) has also appeared that interfaces information from Google Scholar [23]. A social tool named Scholarometer has been developed to facilitate citation analysis and to evaluate the impact of authors [16]. This tool helps to visualize author and discipline networks. There is another tool called SciVal Expert, which visualizes the collaboration and research output of institutions [25]. This tool uses data from Elsevier's Scopus, the largest abstract and citation database of peer-reviewed literature [6]. However, these tools do not provide a visual picture of a single scholar's achievements.

The method and application differ from the prior art. At a glance, Scholar Plot helps the reviewer determine where the researcher's impact (if any) arises from.

Students need more information to decide about their college. Nowadays, a university has a ranking as well as each department with each college in that university. So students need publicly accessible information which is cheap and get a summary of various measures being used to evaluate faculty. Rankings are used to make choices to avoid risks of joining lower ranking colleges [18].

The goal of research is to articulate a clear, comprehensive, and measurable performance evaluation scheme for academics. This scheme should reveal causal relationships among the merit criteria. This research provides a summary interface to facilitate executive decisions. The tool produces a temporal visualization that connects the h-index with the paper citations, and the journal impact factors along with the funding data. Scholar Plot helps the reviewer determine at a glance from where the researcher's impact (if any) arises from.

Here, I introduce a data visualization tool that complements the US News Rankings and the publication information contained in a standard CV. Visualization facilitates access to data and supports actionable insights [26]. It also helps to bring out patterns and pattern violations in the underlying data.

Chapter 3

Methods

In this methods chapter, I will explain various criteria for evaluating academic performance, individual visualization (Scholar Plot) and group visualization (Department Plot) and Academic Garden which is a scalable visualization of academic merit.

3.1 Design Process

There are various criteria for evaluating academic performance. I focus on three main criteria.

- **Impact** - it is the post-production merit. For example, the citations, in which a publication receives. A publication with a higher number of citations has higher visibility. Therefore, I linked the impact to the vertical axis in the plot.

- **Prestige** - this is the pre-production merit associated with the venue of publication. For example, the impact factor of a journal is the merit your publication will acquire because it has been published in that journal. Hence, I associate a disk with variable sizes to the prestige of the venue. I consider it as a 'fancy factor'.

- **Funding** - it enables the production of publications/research. Hence, I placed it at the bottom of the plot. This can help to correlate the production with the funding.

ScholarPlot uses publicly available publication and affiliation information on researchers, scholars, and authors for the purpose of visualizing popular indicators of publishing activity. No single set of indicators can capture all of the dimensions of a publication's scholarly value or an author's contributions to knowledge. Depending on a user's objective, ScholarPlot may be best used in combination with other measures. The visualization consists of a hierarchy of visualization schemes right from the individual to the department and the college.

3.2 Scholar Plot - Individual Visualization

Scholar Plot obtains the Impact Factor (IF) [10] for a particular journal from our database. The data of Impact Factor is acquired from The Thomson Reuters Impact Factor - Web of Science [22]. Based on all this information it constructs the plots as per the design outlined in the Visualization and User Interface section, using nvd3

```
sed -nE '/AwardID|AwardAmount|FirstName|LastName/s/.*>([^<]+)<.*/\1/' *.xml | paste -sd',' -


awk -v ORS="\t" -F '[<>]' '
    /AwardID|AwardAmount|FirstName|LastName/ {print $3}
    FNR == 1 && FILENAME != ARGV[1] {printf "\b\b \n"}
    END {printf "\b\b \n"}
' *.xml > /Users/karlkyeongankwon/Dropbox/scholar-project/NSF-Funding/2014.data


xml sel -t \
    -v //AwardID -o , -v //AwardAmount \
    -m '//Investigator[RoleCode = "Principal Investigator"]' -o , -v FirstName -o , -v LastName  -b \
    -m '//Investigator[RoleCode = "Co-Principal Investigator"]' -o , -v FirstName -o , -v LastName  -b \
    -nl \
*.xml

xml sel -t \
    -v //AwardID -o , -v //AwardAmount \
    -m '//Investigator[RoleCode = "Principal Investigator"]' -o , -v FirstName -o , -v LastName  -b \
    -m '//Investigator[RoleCode = "Former Principal Investigator"]' -o , -v FirstName -o , -v LastName  -b \
    -m '//Investigator[RoleCode = "Co-Principal Investigator"]' -o , -v FirstName -o , -v LastName  -b \
    -nl \
*.xml
```

Figure 3.1: A code snippet of XMLStarlet

reusable charting library [21] and d3.js JavaScript library [3].

The NSF/NIH/NASA funding datasets are available at the respective US government websites in various file formats such as XML, CSV and so on [8, 20]. I implemented a script 3.1 to parse this massive XML dataset into our data structure that consists of AwardID, AwardAmount, First name, Last name, Investigator by RoleCode (Principal Investigator, Co-Principal Investigator and Former Principal Investigator), using XMLStarlet [12]. I imported this data to our database using Toad DBMS tool [24].

Scholar Plot depicts the publications of an individual as a scatter plot and the NSF/NIH/NASA funding as a multiline plot. The publications are represented in a 2D diagram (number of citations vs. year of publication) with the h-index line. The horizontal axis is time, starting with the year of the researcher's first publication

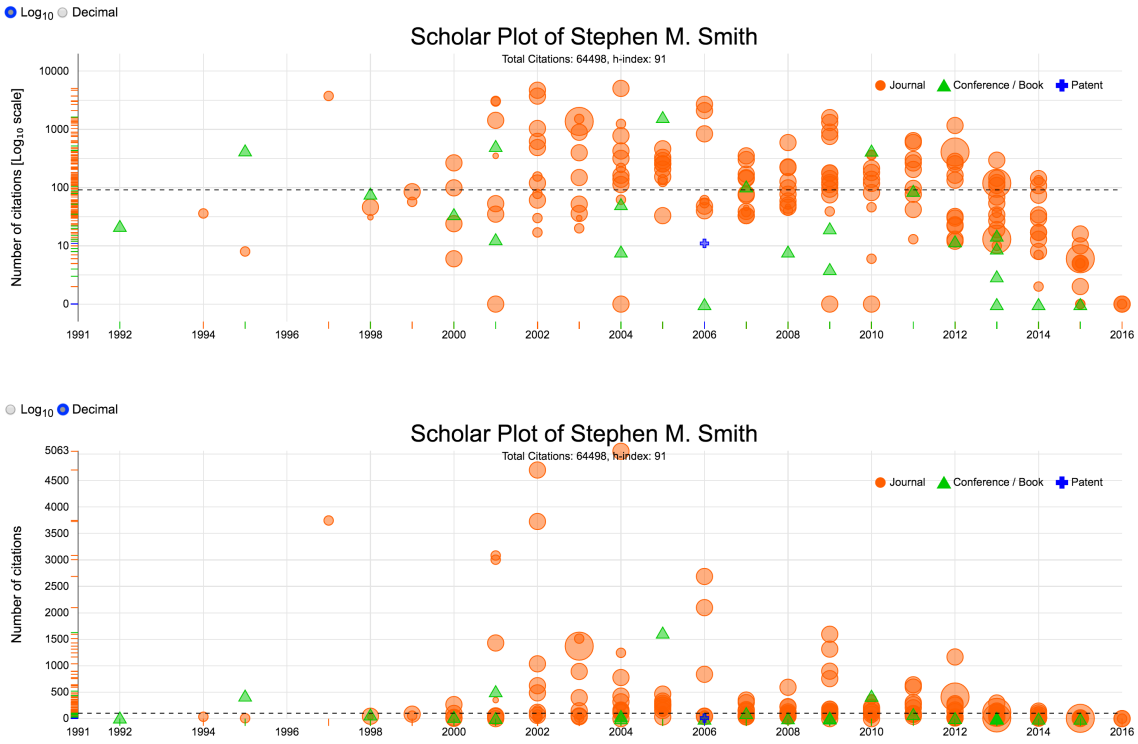

Figure 3.2: An example of a senior records of the log_{10} view and *decimal* view - the radio button allows to switch between different scale views without reloading the entire page. The two different scales view to create a standardized scale for the y-axis for comparison, log_{10} scale is the default plot and an option to toggle to the decimal scale view.

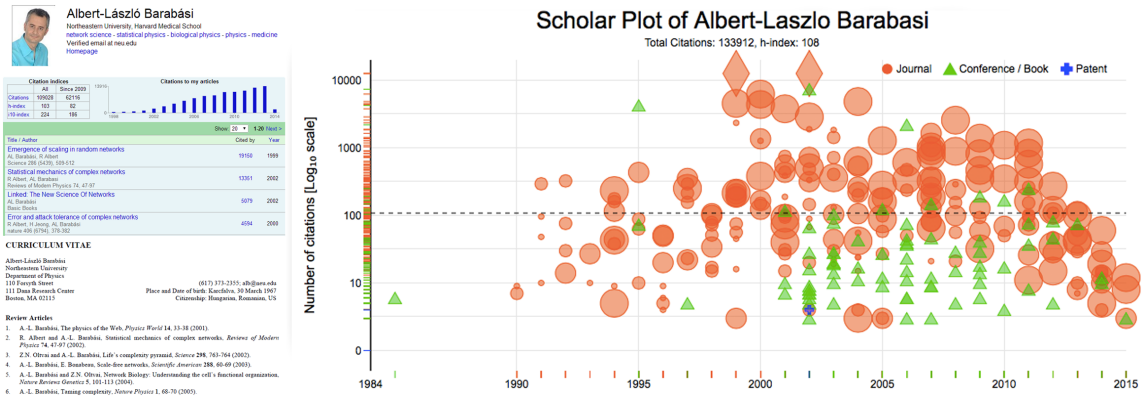

Figure 3.3: An emxaple of a famouse physicist - Google Scholar Profile (Top left) Curriculum Vitae (Bottom left) Scholar Plot (Right) It brings more but simply. Scholar Plot includes all the publications with different colors and symbols, which can lead people to distinguish the type of publication quickly.

and ending with the current year. The vertical axis is the number of citations. The default plot is in log_{10} scale. The user can also view the plot in the decimal scale by a toggle option using a radio button at the top left corner (Figure 3.2). The log scale provides a standardized scale that helps to compare the plots of multiple scholars.

3.2.1 Publication Data

Each publication is represented with a i symbol. The center of the symbol has coordinates (i_{PY}, i_C), where PY stands for Publication Year and C for Number of citations obtained by the publication till date. The journals are represented as circles (orange) with area analogous to the impact factor the journal, and the conferences / books are represented as triangles (green) and the patents as crosses (blue).

By clicking at a symbol, you can obtain the publication title, year, number of

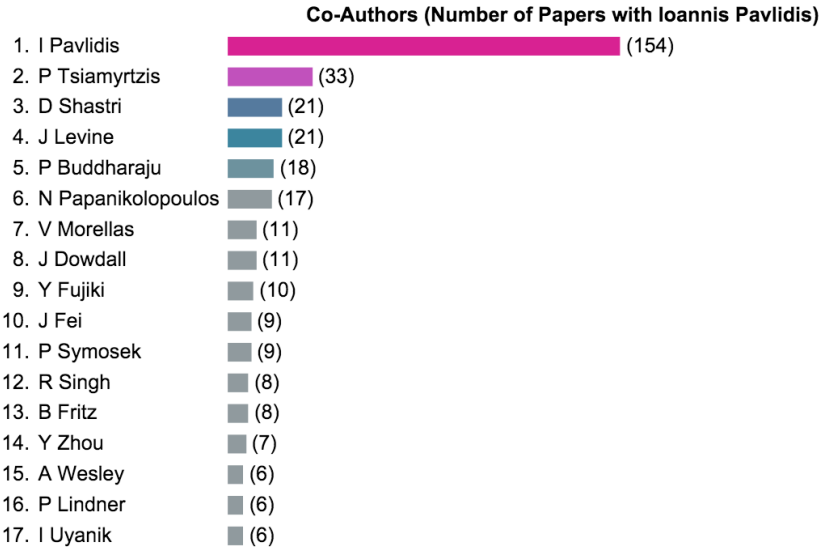

Figure 3.4: The coauthor panel displays the author list.

citations, the venue where published and its impact factor (if it is a journal), as well as the breakdown in the authorship, complete with the level of collaboration between the co-authors and the selected scholar (Figure 3.5). The publication title also enables the user to navigate to the Google Scholar page for the selected paper. This helps to quickly verify and obtain further details of the selected publication. It makes user reach out to the PDF file directly, if available. To improve user experience, I customized the tooltip to give a detailed information without overlapping the plots.

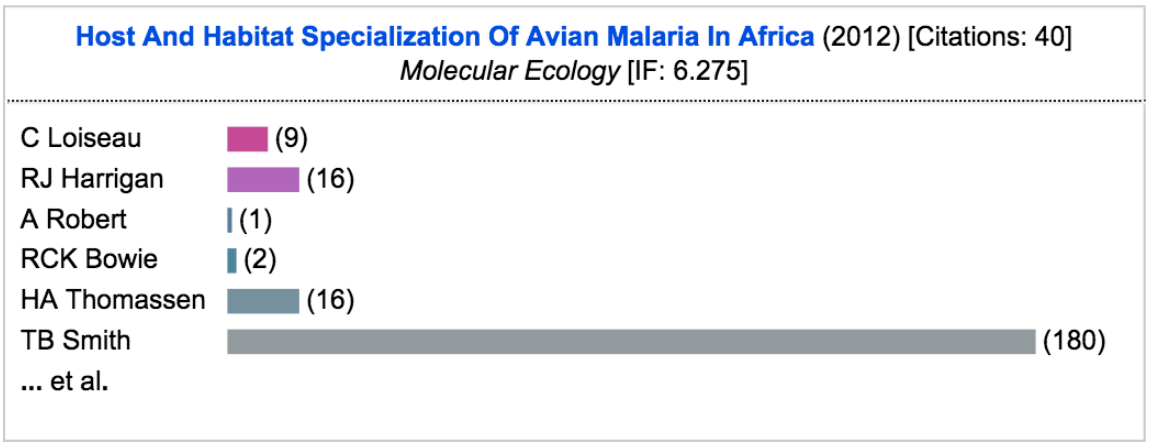

Figure 3.5: An example of the tooltip: the publication title, the year, the number of citations, the venue where published, impact factor, the list of co-authors, the visual horizontal bars with the number of collaboration between the co-authors, and the selected scholar.

The dotted horizontal line on the plot denote the h-index of the scholar. Also, I denote the publications that earned greater than 10,000 citations with diamonds, as they represent the great success in publications (Figure 3.3). The title of the plot contains the name of the scholar and her/his total number of citations along with the h-index. At the top right corner, I distinguishably display a legend shows three different types of publications (Figure 3.6).

I improve user experience to enable users to quickly find and select from a pre-populated list of scholar names as they type. For each character the user enters, scholar plot displays similar matching names on the dropdown list. Even entering

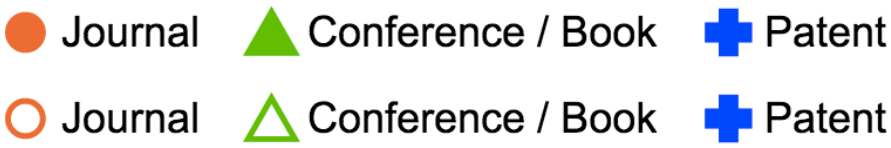

Figure 3.6: The legend allows users to selectively view journals, conferences / books and patents.

the space (“ ”), it displays the 10 most recently inserted scholar’s names. Scholar Plot follows the approach of responsive web design to provide optimal viewing based on the size of screen.

To place the plots in your personal Curriculum Vitae or on a personal web page, I developed the function in server-side, and provided a download button at the top right corner of the plot. This function enables the user to download plots in a zip file. It includes high resolution vector images in SVG (Scalable Vector Graphics) format of the publication and funding plots.

Ranked Density of Publication Types

Scholar Plot also has a projection of the data on the y-axis depicted by small horizontal colored lines. For example, I can clearly see that journals contribute to the h-index of scholar in Figure 3.7 (a) and conferences / books contribute to the h-index of scholar in Figure 3.7 (b). I can conclude the scholar in Figure 3.7 (c)) has many patents and the number of publications within a particular range of citations based on the density of the projected lines.

Scholar plot brings different patterns of scholarly profiles. There are three types

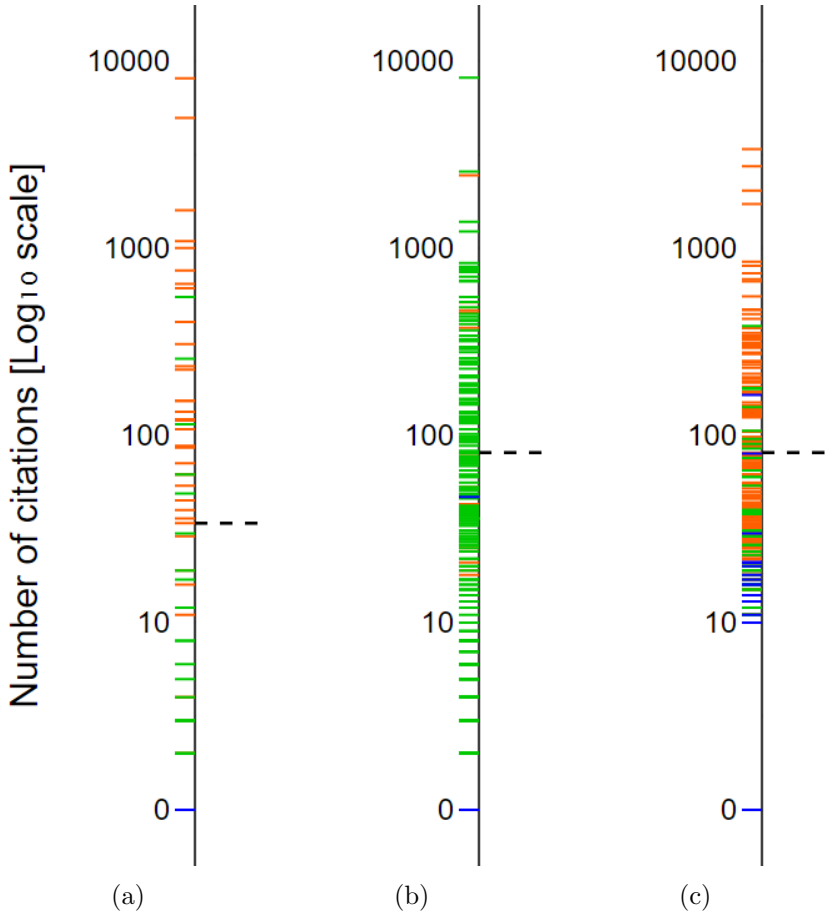

Figure 3.7: Examples of y-axis projection for three different scholars.

of patterns and each examples in shown (Figures 3.8, 3.9, 3.10).

You can bring the journals, patents, and conferences / books in and out of the view by clicking at the respective legend. If there is an overlap between journals, conferences, and patents, this feature can help the user to selectively view them. The user can also zoom into the plot for a closer picture. Also, note that the symbols are not completely opaque. So if there are multiple symbols that overlap, the user can see and interact with them by hovering the mouse over them.

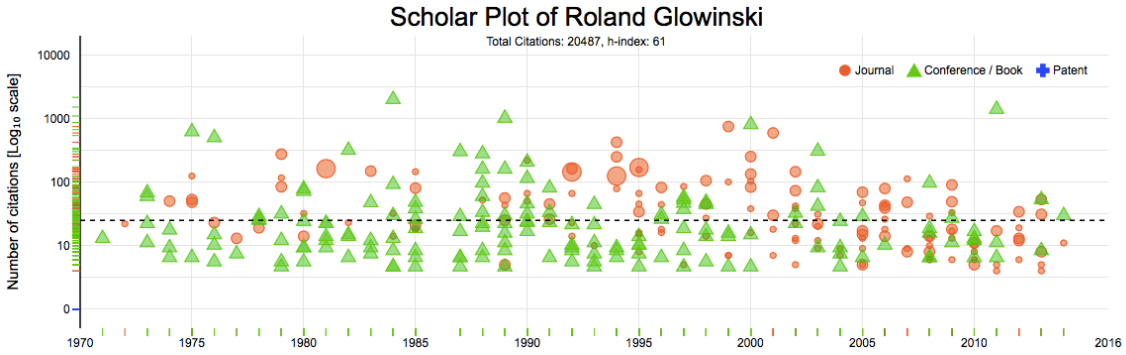

Figure 3.8: Examples of different scholarly profiles - Combination of journal and conference papers

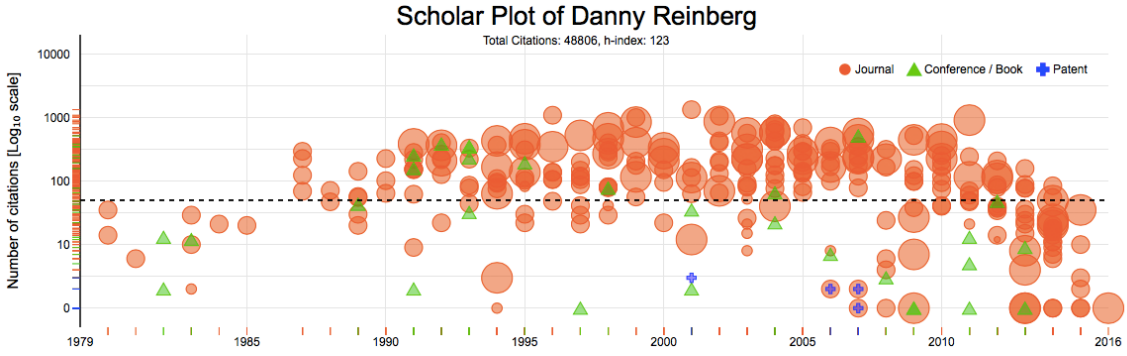

Figure 3.9: Examples of different scholarly profiles - Preponderance of journal papers

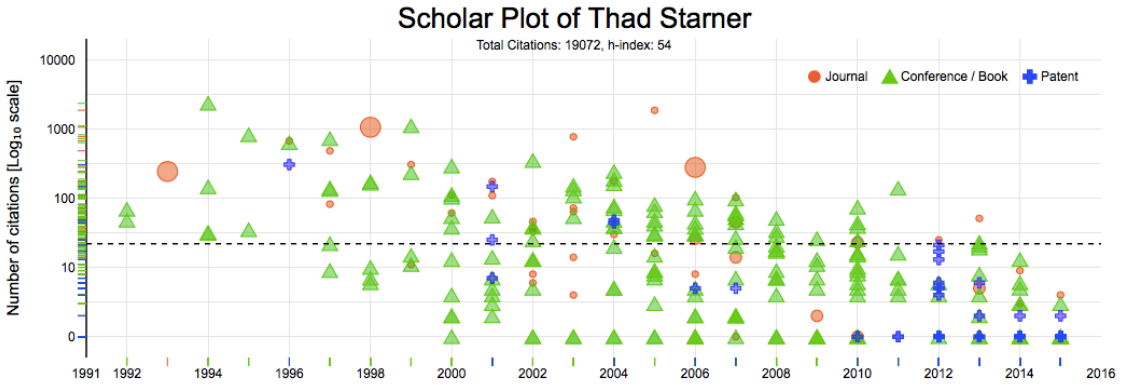

Figure 3.10: Examples of different scholarly profiles - Combination of conference papers and patents

Disk Size - How to determine the size of disks

I wanted to plot a more efficient visual for Journal publications, which is presented by different sizes, that tells the ranking of Journal by Impact Factor Index. To do this, I analyzed the data set of JCR 2015 IF and ran a quartile function as a useful concept in statistics to determine the size of disks in Scholar Plot. Based on this number, the system will decide the size of plot of each journal data and plot it in real-time 3.12. The quartile values are shown in Figure 3.11.

3.2.2 Funding Data

Scholar Plot also depicts the NSF/NIH/NASA funding of an individual as a multi-line (Figure 3.13). Each breakpoint in the multiline corresponds to the individual's total amount in all NSF/NIH/NASA awards for the specific year. By pointing at a

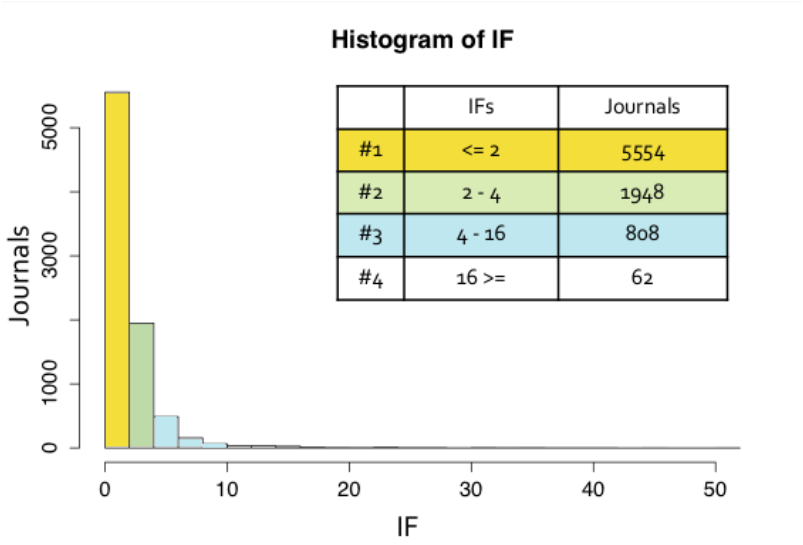

Figure 3.11: Histogram of Impact Factor Jounal

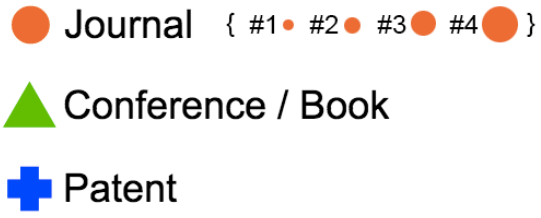

Figure 3.12: Disk size along with journal

breakpoint you can obtain the NSF/NIH/NASA awards IDs, award amounts, and the investigator's role. The total annual funding information per year is also available by clicking the legend.

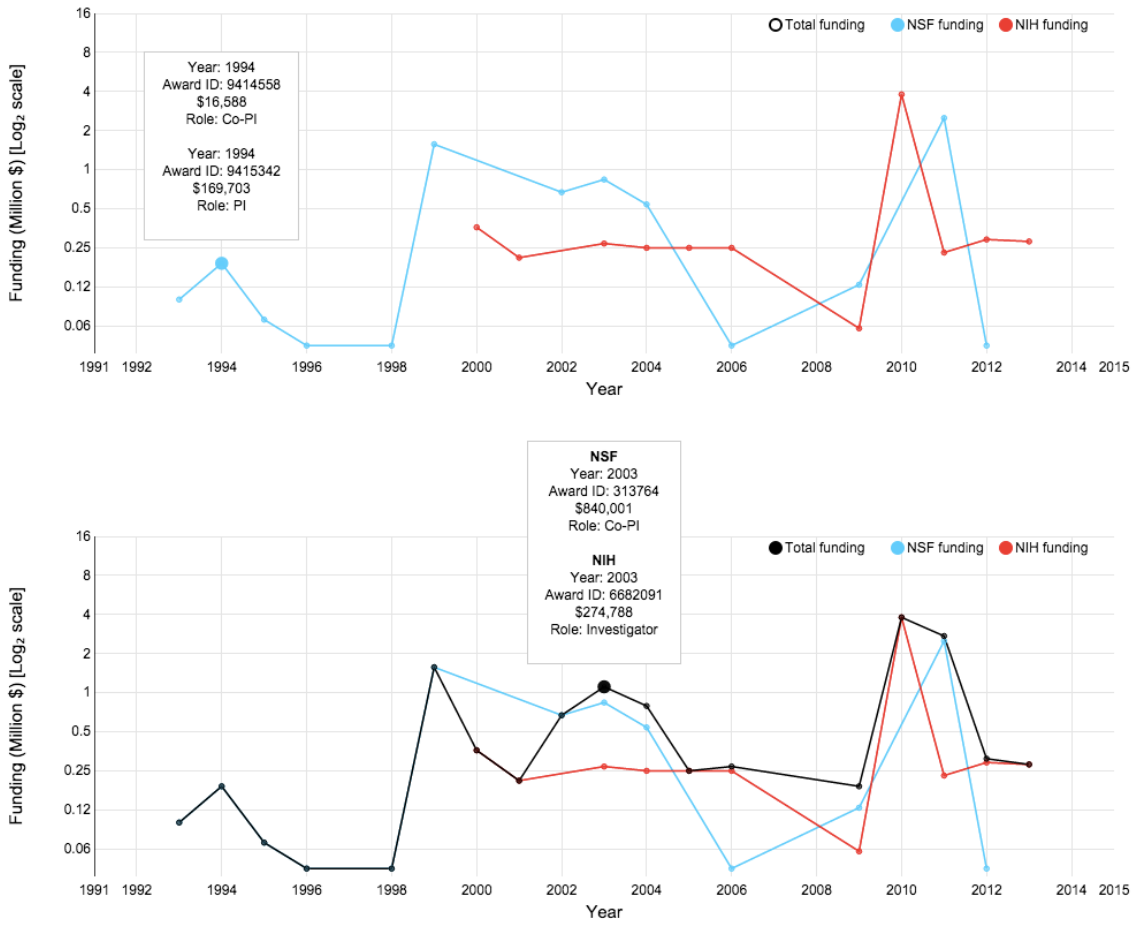

Figure 3.13: An example of Scholar Plot - Visualizing Funding Data

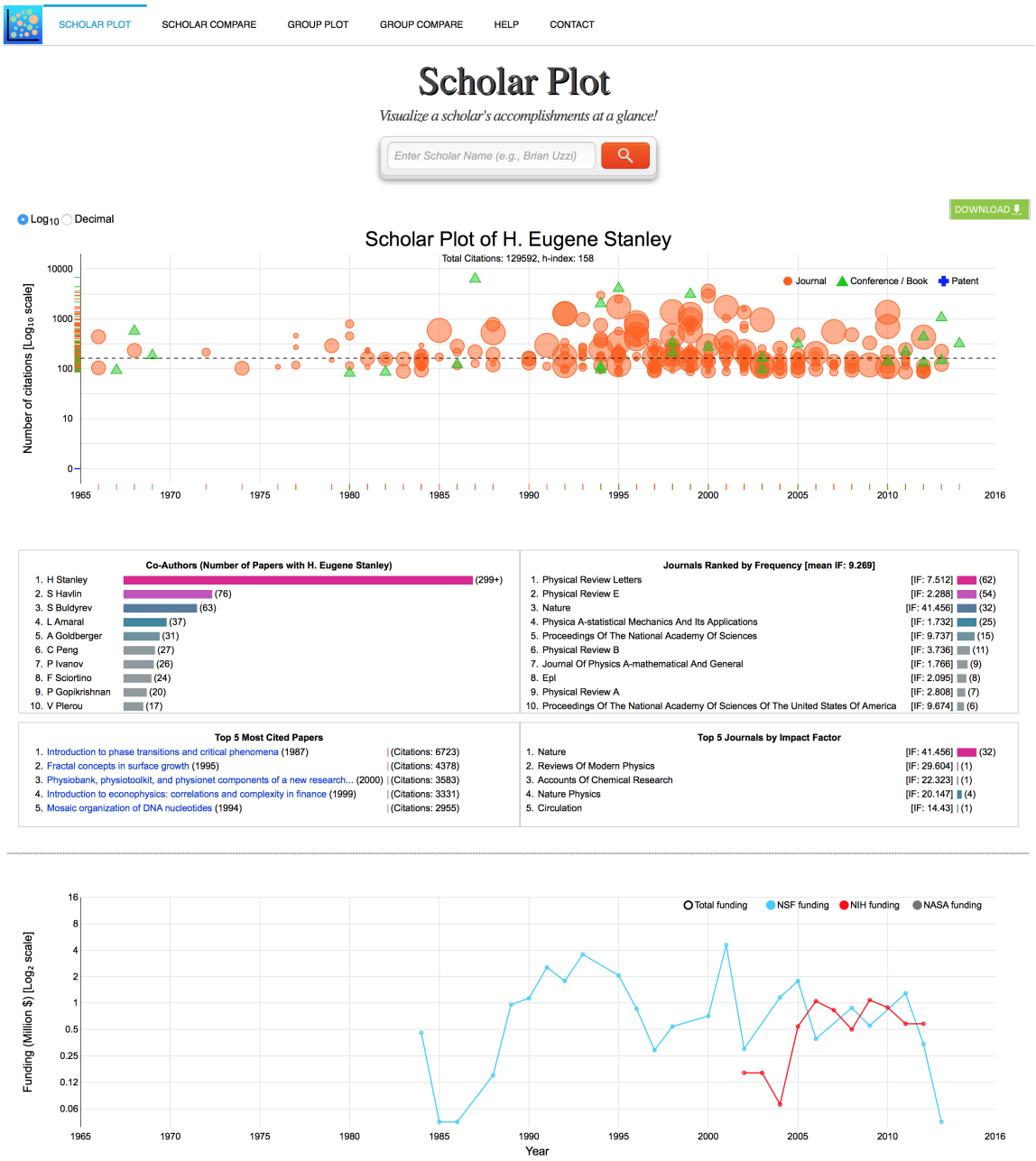

Figure 3.14: Base level Scholar Plot (SP) example - a famous physicist and interdisciplinary scientist with dozens of articles in *Nature*. The summary panels in the middle were added after a feedback from the focus group. Notice how this scholar's publication production exploded in sync with the commencement of substantial federal funding.

3.3 Department Plot - Group Visualization

The group level of Scholar Plot visualizes department/college academic records. One of the important issues was to determine how to scale the individual visualization to the group level. Group plot consists of 2 aspects - plot at the department level and at the college level. I applied our design philosophy at the group level. I use pie charts and bar charts (Figure 3.15) to display the information in a compact manner. Pie charts are useful to show a proportion of contribution of each individual to the group (i.e. department). For pie-charts, I displayed the top 5 scholars to avoid overcrowding the pie chart (Figure 3.24).

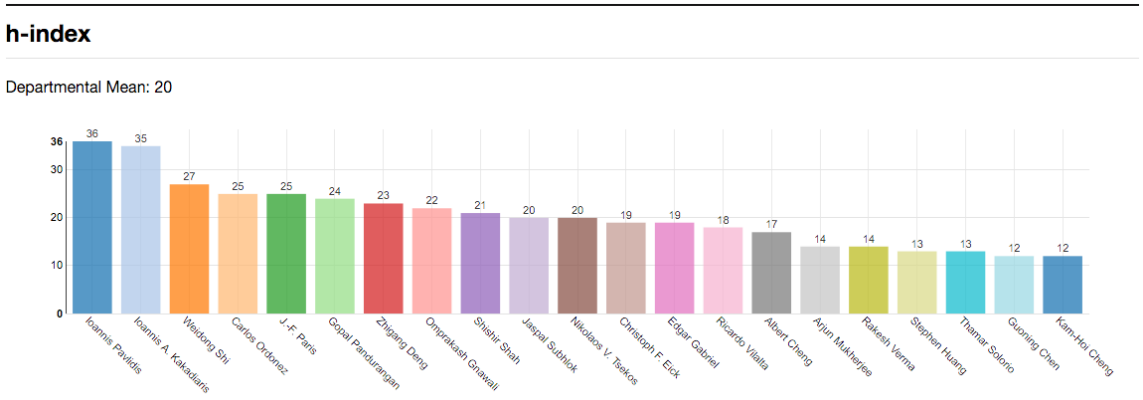

Figure 3.15: Example of bar chart by h-index.

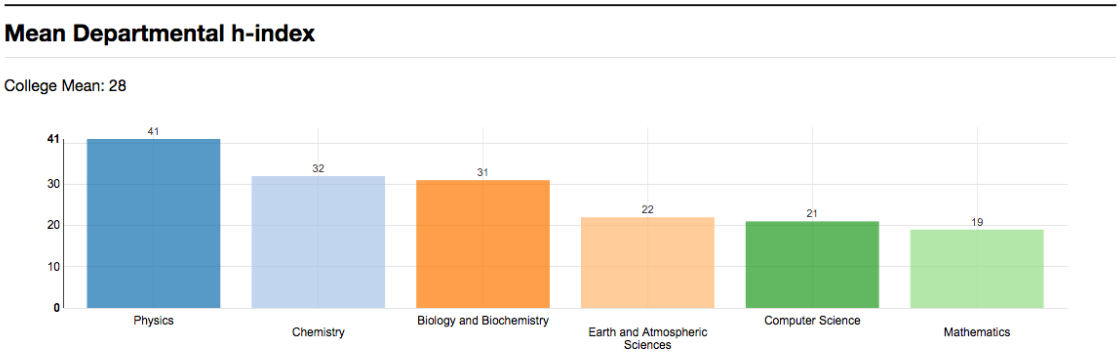

Figure 3.16: Mean Departmental h-index - College of Natural Sciences and Mathematics at University of Houston

3.3.1 Department Plot

Departmental Plot is an attempt to visualize aspects of tenured and tenure-track faculty contributions to their home departments. These aspects are not only intellectual contributions and perhaps not even the most important. The faculty are compared based on publicly available measures like h-index, citations and impact factor. I visualized a citation contribution as a pie chart normalized by the number of years in which a scholar spent in academia. Also, I portrayed charts depicting the highest (Home Run) cited paper and the highest (Home Run) impact factor journal where the scholar published.

Figure 3.17: Home Run Citations, Mean Departmental Citations, Citations Pie Chart (Total Citations and Normalized Citations) - College of Natural Sciences and Mathematics at University of Houston

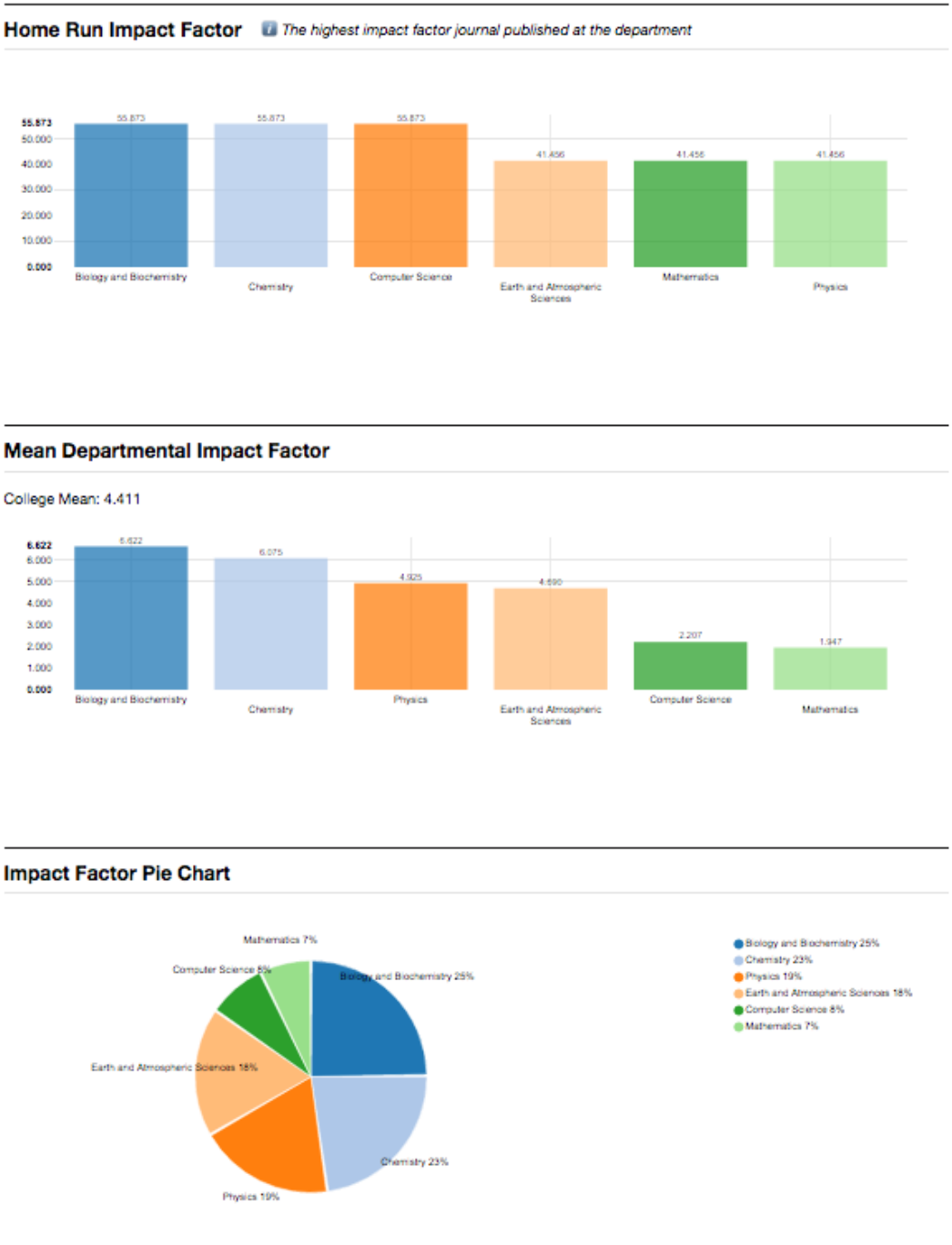

Figure 3.18: Home Run Impact Factor, Mean Departmental Impact Factor, Impact Factor Pie Chart - College of Natural Sciences and Mathematics at University of Houston

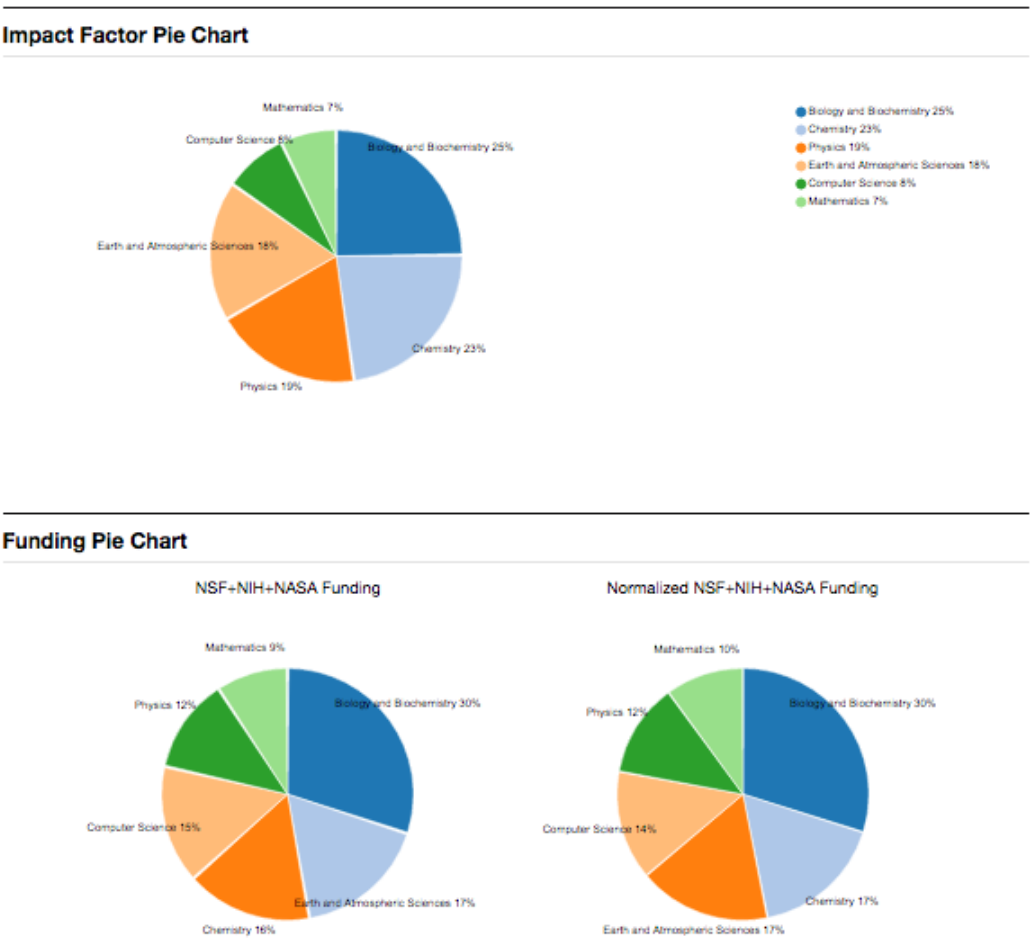

Figure 3.19: Funding Pie Chart - College of Natural Sciences and Mathematics at University of Houston

3.3.2 Department Plot Glossary

H-INDEX

The h-index is a form of measure that takes into account the number of citations and number of total publications made by a scholar.

HOME RUN CITATIONS

The Home Run Citations bar chart shows the highest cited paper within the department, with the largest number of citations of a publication on the y-axis, and the scholar associated with that publication on the x-axis.

CITATIONS

The Citations bar chart displays the total amount of citations a scholar has received through all of his/her publications.

CITATIONS PIE CHART

The Citations Pie Chart displays the percentage of citations that each scholar produced out of all the citations in the department.

HOME RUN IMPACT FACTOR

The Home Run Impact Factor bar chart shows the highest journal impact factor for each of the deparment's scholars. The impact factor is on the y-axis and the name of the scholar is on the x-axis.

MEAN IMPACT FACTOR

The Mean Impact Factor bar chart shows the average level of journal impact factor by each scholar in the department with the impact factor on the y-axis

and the name of the scholar on the x-axis.

IMPACT FACTOR PIE CHART

The Impact Factor Pie Chart displays the percentage of journal impact that each scholar is responsible for out of the total journal impact of the college.

FUNDING

The funding bar chart shows the amount of funding awarded to each scholars in the department with the amount of dollars in funding on the y-axis and the scholar associated with that funding on the x-axis.

FUNDING PIE CHART

The Funding Pie Chart displays the percentage of funding that each scholar has received in the department.

3.3.3 College Plot

College plot attempts to visualize the contributions of the departments to the home college. College plot pictures the mean values of various measures described above for each department. I used pie charts (Figure 3.24) and bar charts like in the department plot. Note that the data for department and college plot is generated by using a query to our database.

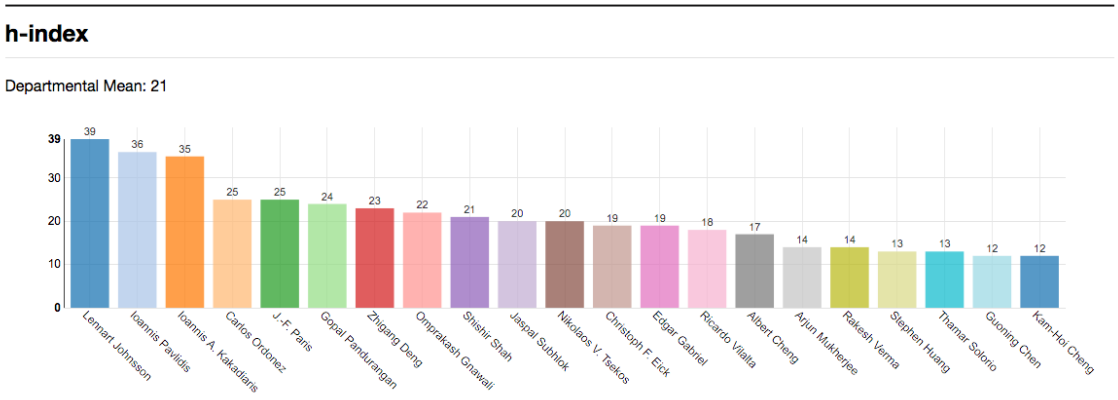

Figure 3.20: hIndex - Department of Computer Science at University of Houston

3.3.4 College Plot Glossary

FUNDING PIE CHART

The Funding Pie Chart displays the percentage of funding that each scholar has received in the department.

MEAN DEPARTMENTAL H-INDEX

The Mean Departmental h-index bar chart shows the average h-index for each department in the selected college. The y-axis is the mean h-index and the x-axis is the department name.

HOME RUN CITATIONS

The Home Run Citations bar chart shows the number of citations of the highest cited paper within each department on the y-axis and the names of the departments on the x-axis.

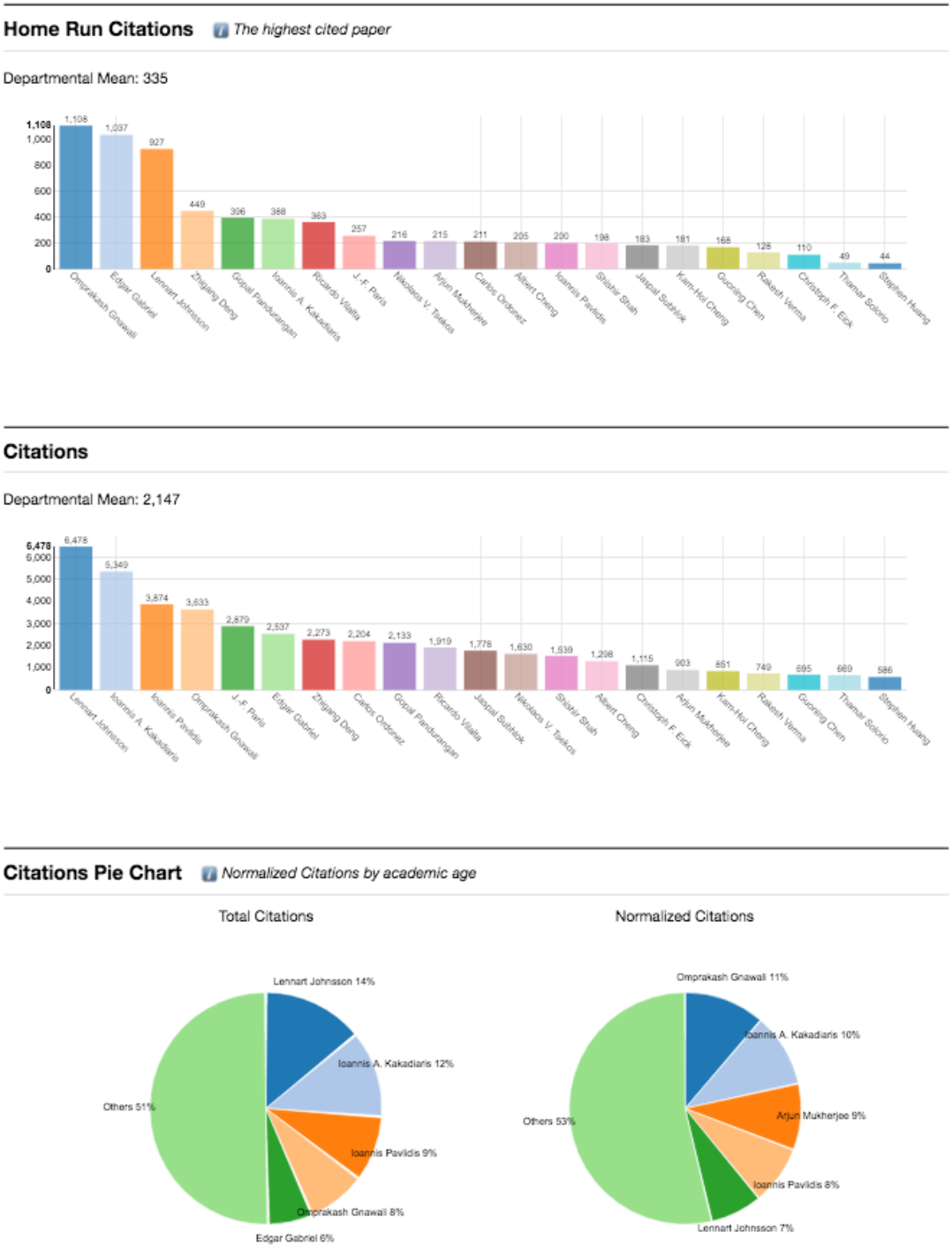

Figure 3.21: Home Run Citations, Citations, Citations Pie Chart (Total Citations, Normalized Citations) - Department of Computer Science at University of Houston - Citation

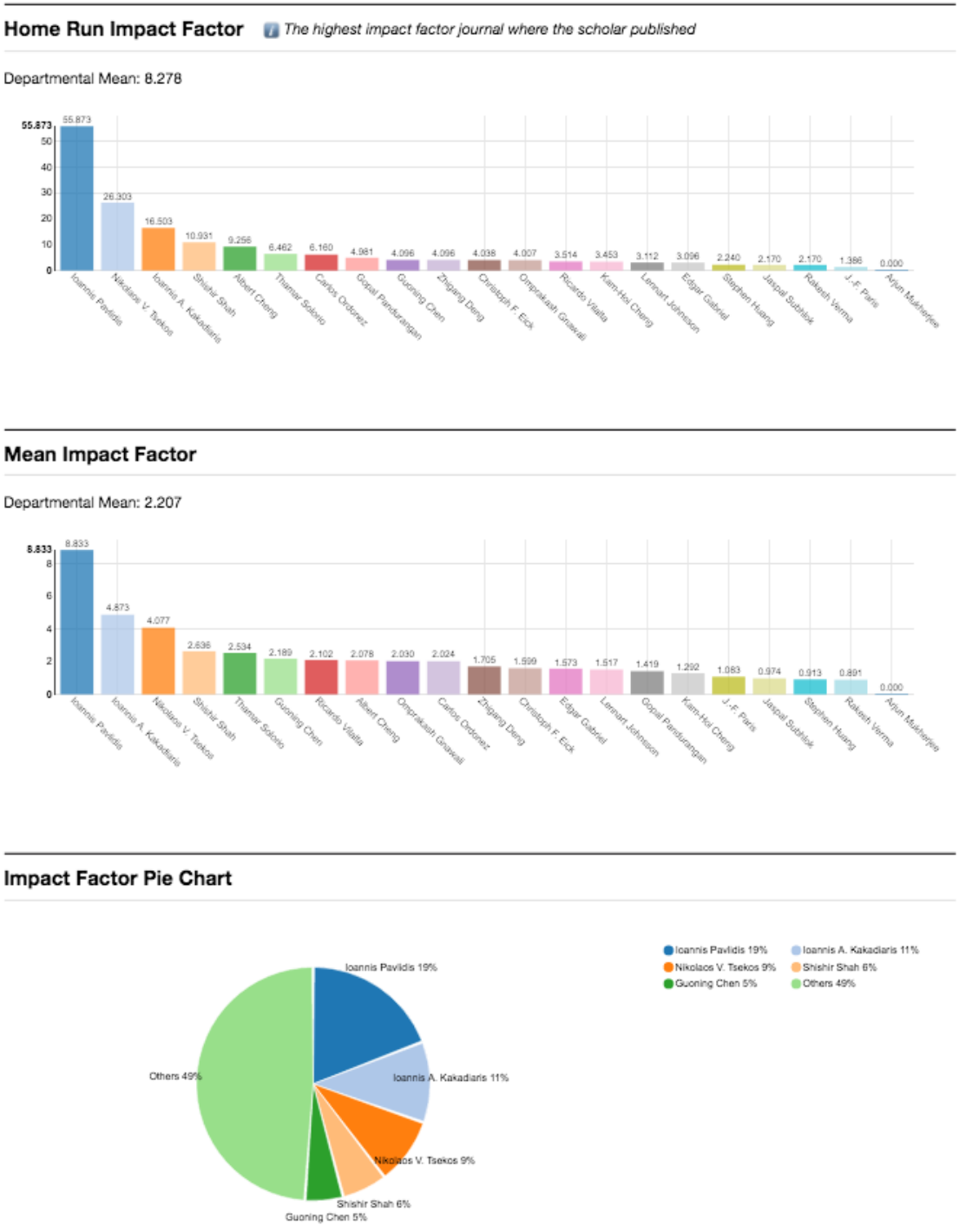

Figure 3.22: Home Run Impact Factor, Mean Impact Factor, Impact Factor Pie Chart - Department of Computer Science at University of Houston - Impact Factor

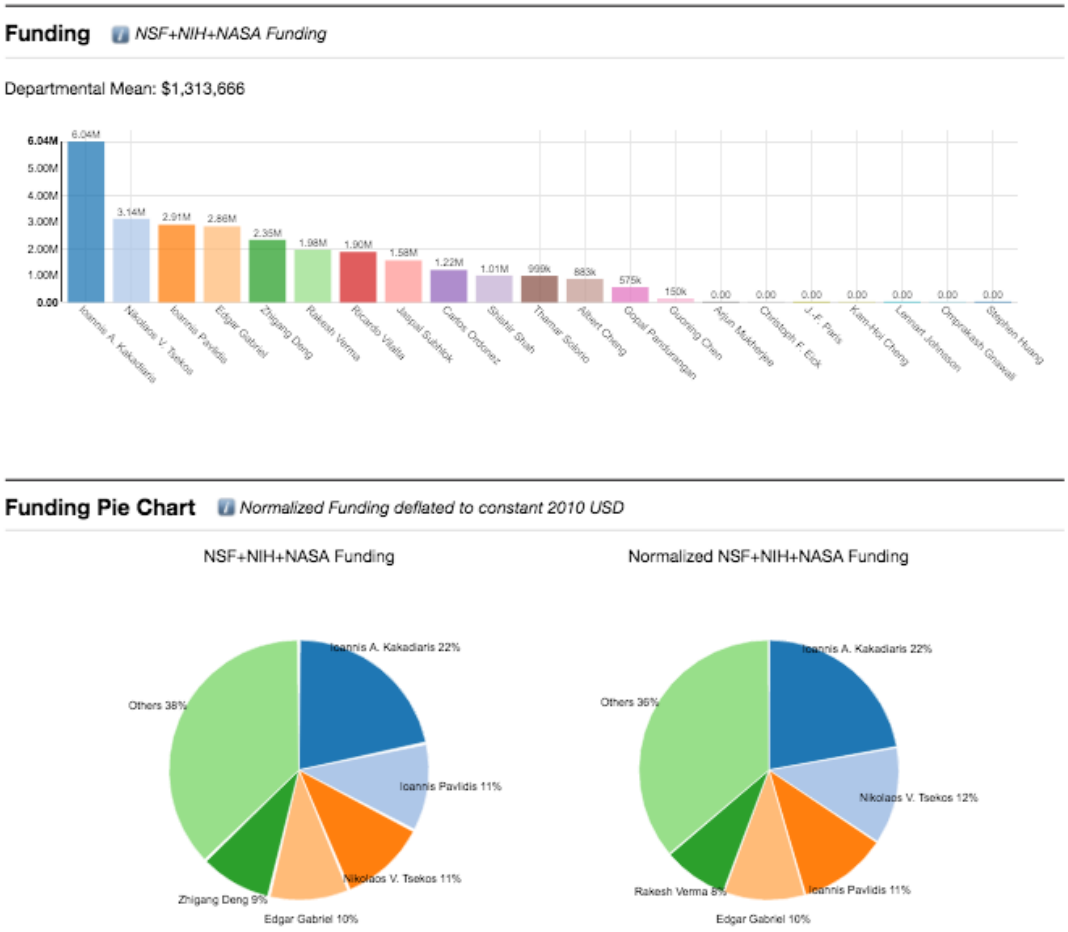

Figure 3.23: Funding, Funding Pie Chart (NSF+NIH+NASA Funding, Normalized NSF+NIH+NASA Funding) - Department of Computer Science at University of Houston - Funding

MEAN DEPARTMENTAL CITATIONS

The Mean Departmental Citations bar chart shows the average number of citations for the scholars in each department with the number of citations on the y-axis and the names of the departments on the x-axis.

CITATIONS PIE CHART

The Home Run Impact Factor bar chart shows the highest journal impact factor of each department with the impact factor on the y-axis and the names of the departments on the x-axis.

HOME RUN IMPACT FACTOR

The Citations Pie Chart displays the percentage of citations that each department produced out of all the citations in the college.

MEAN DEPARTMENTAL IMPACT FACTOR

The Mean Department Impact Factor bar chart shows the average level of journal impact factor by each department with the impact factor on the y-axis and the department on the x-axis.

IMPACT FACTOR PIE CHART

The Impact Factor Pie Chart displays the percentage of journal impact that each department is responsible for out of the total journal impact of the college.

FUNDING PIE CHART

The Funding Pie Chart displays the percentage of funding that each department has received in the college.

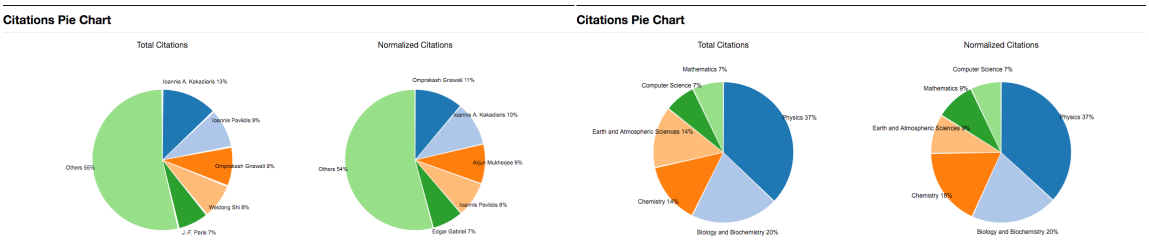

Figure 3.24: Example of Citations Pie Chart. The ones on the left are at the Department Level, the ones on right are at the College Level. The charts depict total citations and normalized citations.

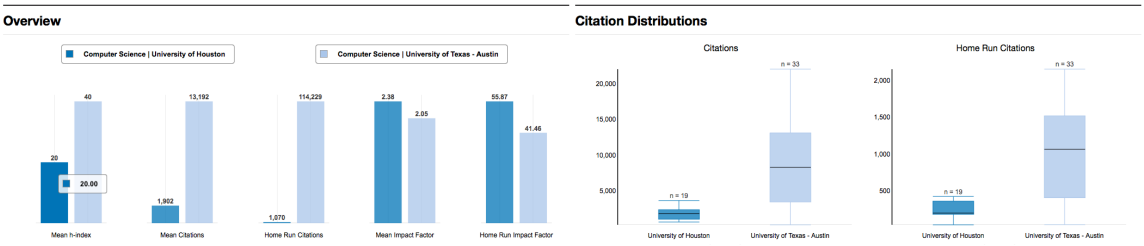

Figure 3.25: Example of Group Compare between Departments of Computer Science at University of Houston and the University of Texas - Austin.

3.3.5 Compare Plot

Department Compare aims to assist people to have a deeper understanding of the inner accomplishments in the departments. It complements the ranking given to the department by the US News Report. Department Compare feature compares the departments with the same publicly available measures. Scholar plot compares the summary statistics like the mean values. It uses box plots to compare the distribution of values of the individual faculty in each department. To determine whether a result is statistically significant, box plots denote the significant sign.

3.3.6 Compare Plot Glossary

OVERVIEW

The overview displays the mean h-index, mean citations, and home run citations for the two departments in the form of multiple bar graphs. Each color represents one of the departments, indicated by the key at the top of the section.

H-INDEX DISTRIBUTIONS

The h-index distributions chart displays the spread of the h-indexes within a department through box plots. The n value above each box plot represents the sample size used. The stars below the title of each section signify the statistical significance.

CITATION DISTRIBUTIONS

The Citations Distributions chart shows the spread of the citations within the department for each university. The n value above each box plot represents the sample size used. The stars below the title of each section signify the statistical significance.

IMPACT FACTOR DISTRIBUTIONS

The Impact Factor Distributions chart displays the spread of citations within a department through box plots. The n value above each box plot represents the sample size used. The stars below the title of each section signify the statistical significance.

3.4 Academic Garden - Scalable Visualization

Academic Garden (AG) is a scalable visualization of academic merit. It applies to individual academics, departments, colleges, and any other academic group thereof, such as a research lab or a project team. Reminiscent of the legal views for physical personhood and corporate personhood, we consider that individual academics and academic groups share behavioral characteristics. Specifically, we argue that academic performance has three pillars that are scale invariant: (a) funding that enables intellectual production; (b) prestige of the venues where intellectual products appear; and, (c) impact of the intellectual products. In the case of groups, these three variables are expressed as statistics of the corresponding individual measurements.

Academic Garden uses the flower metaphor to visually articulate performance for academic entities. The width of the flower's stem is commensurate to the academic funding this entity received ('juice conduit'). The height of the flower's stem is commensurate to the impact of the entity's intellectual products ('visibility'). The diameter of the flower's disc is commensurate to the prestige of the venues where these products appeared ('fancy factor').

3.4.1 Research Funding: Enabler of Production

Research funding is an enabler of academic production. Very few things can be done in the absence of funding in science and engineering. Even in humanities, some funding is needed in many cases (e.g., travel support for archival research). Research funding is dispensed through peer-reviewed proposal competitions, and

for this reason, it is not only an enabler but also has inherent merit. As different disciplines need different levels of funding some normalization is in order. This normalization can be any statistic. We prefer the quartile where the funding level of the academic entity's record belongs with respect to all the records in the specific discipline. 'All' here is commensurate to the selected reference, whether this is a university department or a set of departments across the United States. Needless to say that the original funding records need to be adjusted, taking into account the entity's age (if the entity is a physical person) or the number of individuals participating in the entity's personhood (if the entity is a group).

3.4.2 Prestige of Product Venue: Pre-production Achievement

Funded (and unfunded) research typically results in intellectual products. These are typically journal papers, conference proceedings papers, or books. Occasionally, intellectual products include patents or software packages, such as smartphone applications. Almost all intellectual products undergo review process, and the ones successfully passing this review process have inherent merit. The review process criteria are not uniform. Moreover, publishing in different venues is associated with various degrees of difficulty. In journals, this difficulty is largely associated with the journal's impact factor (IF), as determined by Thomson Reuters - the higher the IF, the more difficult it is to be published in a journal, and the more valuable and prestigious a potential acceptance. For refereed conferences, the prestige is loosely

associated with the venue's acceptance rate? the lower the acceptance rate, the more difficult it is to get into the conference proceedings, and the more prestigious the accomplishment. Unfortunately, there is no universally accepted ranking list for conferences, as is the case of the Thomson Reuters IF list for journals. Hence, it is not opportune to assign a numeric score to conference publications. The same applies for books, where evaluations are even more qualitative, and based on opinions about the perceived prestige of the publishing house. And, we are totally agnostic regarding pre-production credit, when it comes to patents and software products. As a result, for the moment we use only IF to measure pre-production achievement. Based on the histogram analysis of the frequency of publications in the IF list of journals, we use four classes to group prestige. Different grouping may be adopted, however, depending on the analytics used.

- CLASS-1: IF $<$ 2
- CLASS-2: 2 $<=$ IF $<$ 4
- CLASS-3: 4 $<=$ IF $<$ 16
- CLASS-4: 16 $<$ IF

3.4.3 Product Impact: Post-production Achievement

Once a paper appears in a journal or conference proceedings, or a book appears in the market, it gets noticed and depending on how useful researchers find the concept or method contained therein, they may start using it, and citing its source in their

own intellectual products. This practice constitutes impact, which is a sought-after outcome of the research process as the building block of scientific advances. There are several ways of measuring impact, but the most widely accepted is the citation count.

As different disciplines have different population sizes and publication practices, which may affect citation numbers, normalization is in order. This normalization can be any statistics. We prefer the quartile where the citation count of the academic entity's record belongs with respect to 'all' the records in the specific discipline. 'All' here is commensurate to the selected reference, whether this is a university department or a set of departments across the United States. Needless to say that the original citation records need to be adjusted, taking into account the entity's age (if the entity is a physical person) or the number of individuals participating in the entity's personhood (if the entity is a group).

Putting it All Together

The design is not only measurable, but also comprehensive, fair, and sensible. As an abstracted pattern, it holds true not only for academic production, but also for many other types of creative production. I considered representative cases to support the argument that this value system gives credit where credit is due, while at the same time it pinpoints the hidden truth that are not accounted for under the present heuristic and fuzzy evaluation processes.

SINKHOLE: Take the case of a well-funded academic entity that churns out

products appearing in low-level journals and collecting few citation hits. This entity deserves some credit for winning competitive grants. From the science policy point of view such an entity is a liability in the long run, as it acts like a sinkhole of public funds. The three-prong merit system captures the pros and cons of this case, highlighting their causal linking.

LEAN & MEAN: In contradistinction, consider an entity that has moderate funding but publishes articles in highly prestigious journals that receive many citation hits. From the science policy point of view, this is a 'lean and mean' academic machine, as with moderate resources achieves maximum results. Every relevant funding agency would like to give to this entity more funding, as it represents a great investment. The three-prong merit system captures the pros of this case, highlighting their causal linking.

ODDBALL: Consider an entity that publishes highly novel concepts in big journals. The concepts attract attention for their creative power but find no use for the moment, receiving few citation hits. The fact that the concepts did not find an immediate application does not detract from their intellectual worth, which is captured by the pre-production merit criterion. The three-prong merit system captures the pros and cons of this case, highlighting their causal linking.

UNASSUMING HERO: Consider an entity that publishes specialized methods in solid transaction level journals. These methods find wide applicability in the relevant disciplinary communities and are widely cited. This entity did not receive any huge pre-production merit. However, its post-production impact more made up for it. The three-prong merit system captures the pros and cons of this case, highlighting

their causal linking.

3.4.4 Academic Garden Flower Diagram

The flower was chosen as the visual metaphor for the performance of an academic entity. A nice looking flower is highly desirable, and so is a meritorious academic entity. Structurally, the stem, and disc make up a flower. We defined: (a) the width of the stem to be commensurate to the academic entity's funding; (b) the height of the stem to be commensurate to the academic entity's citation record; and (c) the diameter of its disc to be commensurate to the prestige of the venues where it publishes.

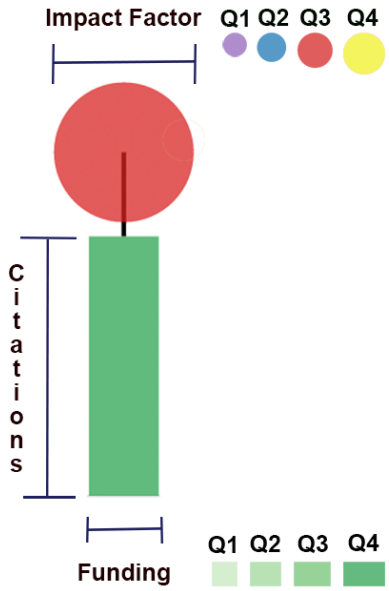

Figure 3.26: A wider stem means that a flower has the necessary support to grow. The width of each stem in the plot indicates the level of funding the scholar has received. A higher quartile of funding is represented by a wider stem and a darker green color. As a flower grows its stem heightens. The length of each stem in the plot represents a scholar's total number of citations.

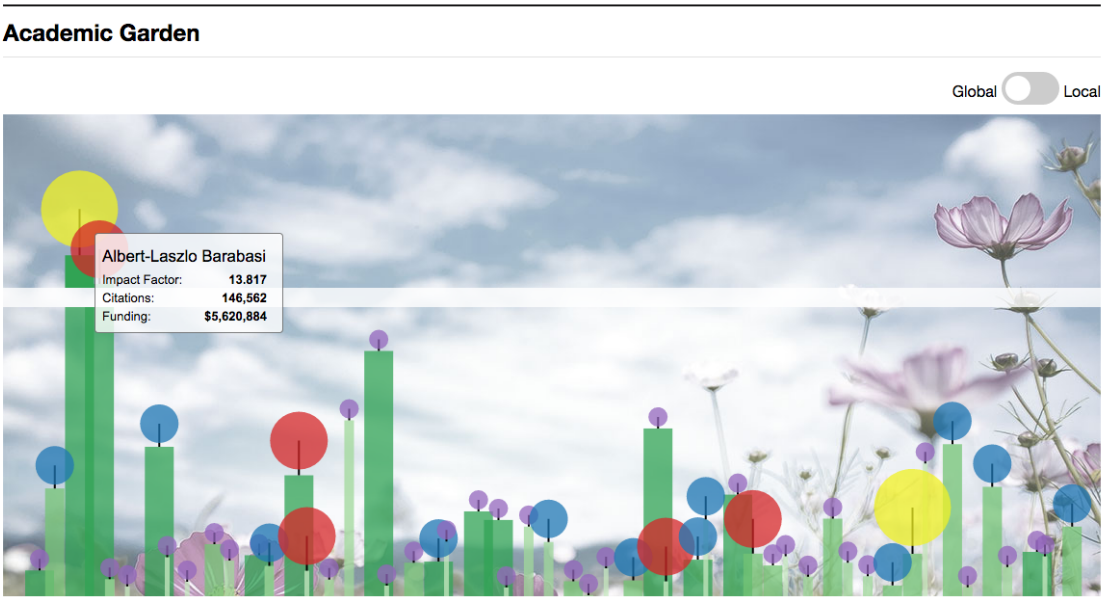

Figure 3.27: Academic Garden example of Global Scale - Computer and Information Science at Northeastern University.

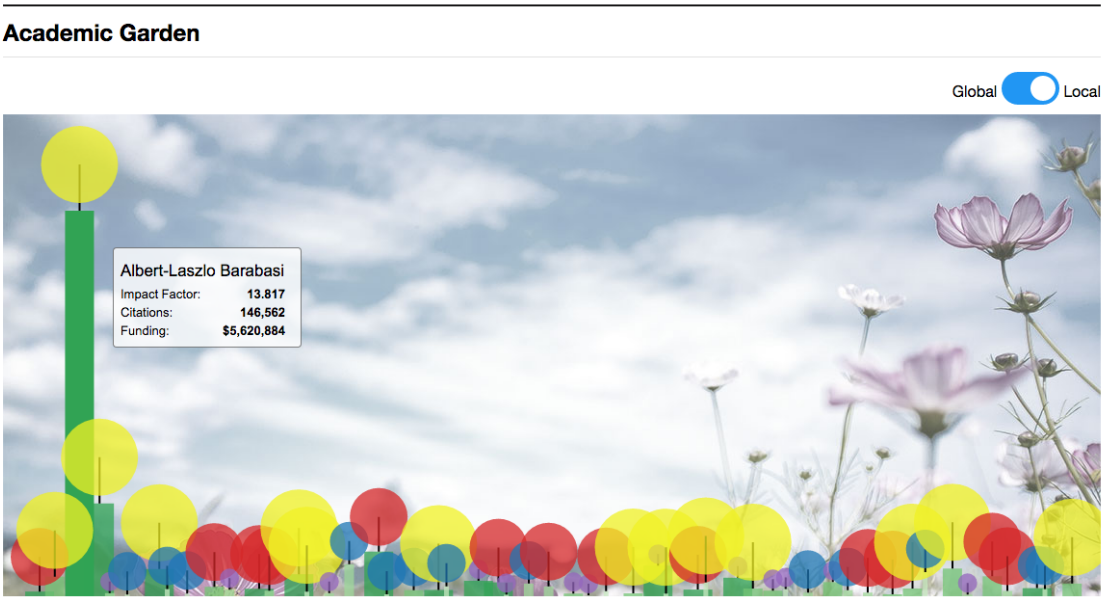

Figure 3.28: Academic Garden example of Local Scale - Computer and Information Science at Northeastern University

Chapter 4

Software Design

4.1 Data Sources

There were several options to get bibliographic data for powering the publication plot of Scholar Plot (SP). These include Scopus, ISI Web of Knowledge, and Google Scholar. We chose Google Scholar for two reasons: a) it is all inclusive, covering all types of publications such as journals, conferences, books, and patents; and, b) it is freely available. Scopus is subscription based and not as inclusive as Google Scholar. ORCID has publications and funding data but requires extensive set-up.

Our choice carries a few challenges, too. Google Scholar does not provide an application programming interface. Hence, we had to develop an elaborate software to scrape information off publicly available Google Scholar pages. Also, not every academic has a Google Scholar page.

We use the Journal IF List that is issued every year by Thompson Reuters to assign disk sizes to journal publications.

For funding records, we use the publicly available grant records from the National Science Foundation (NSF) [8], the National Institutes of Health (NIH) [20], and the National Aeronautics and Space Administration (NASA) [1]. These are the only funding agencies with publicly available datasets at this point.

Agencies	Fiscal Year	Rows	Per Year
NSF	FY 1985 - FY 2013	312,311 rows	10,769/year
NIH	FY 2000 - FY 2013	777,657 rows	55,456/year
NASA	FY 2007 - FY 2015	16,670 rows	1,852/year

Table 4.1: Funding datasets in Scholar Plot system.

4.2 System Architecture

Scholar Plot is the web-based data visualization method that uses HTML5, CSS3, and SVG to render a scholar's accomplishment at a glance. We created a MySQL database to store the mapping between the scholar names and their Google Scholar IDs. We also designed and created database tables for NSF/NIH/NASA funding data. The user can search the name of the scholar in a text field. When the user starts to enter the name of the scholar, the names in our database which are similar to the entered name will be listed as a drop down list. We use jQuery and Ajax (asynchronous JavaScript and XML) method to have this feature, which connects to the database to get the list of names. If there are no matching/similar names, the user can also insert her/his Google Scholar ID to the database by one click event.

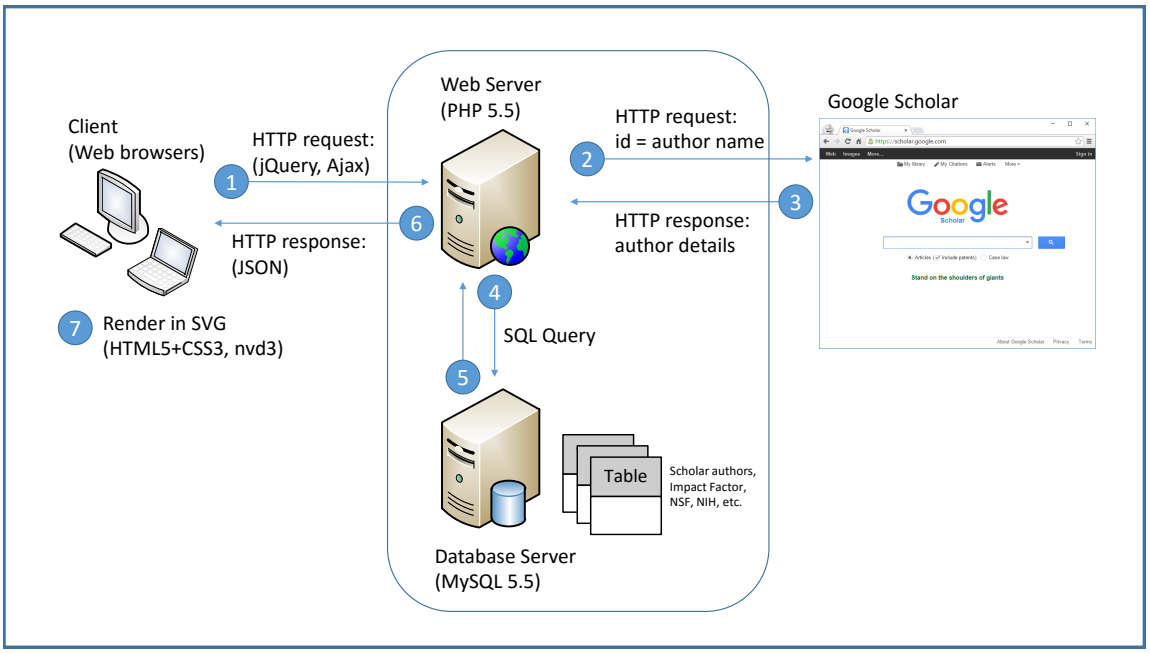

Figure 4.1: System Architecture of Scholar Plot.

Once the scholar's name is selected, the user can run the application to see the visual results of the selected scholar's publications and fundings. Scholar Plot connects to the Web server to retrieve the necessary information. The server-side application is implemented in PHP scripting language and MySQL. The HTTP protocol is used for communicating between client-side and server-side to get the basic information via JSON format (JavaScript Object Notation) and JSONP function (Figure 4.1). Scholar Plot also uses htmlSQL library to parse Google Scholar's page to extract user basic information [14].

Scholar Plot obtains the Impact Factor (IF) for a particular journal from our

database. The data of Impact Factor is acquired from The Thomson Reuters Impact Factor - Web of Science. Based on all this information it constructs the plots as per the design outlined in the Visualization and User Interface section, using nvd3 library [21].

The NSF/NIH/NASA funding datasets are available at the respective US government websites in various file formats such as XML, CSV and so on [8, 20, 1]. We implemented a script to parse this massive XML dataset into our data structure that consists of AwardID, AwardAmount, First name, Last name, and Investigator by RoleCode (Principal Investigator, Co-Principal Investigator, and Former Principal Investigator), using XMLStarlet [12]. We imported this data to our database using Toad DBMS tool. Currently, we have only these three funding data sources. So this is a limitation of the current system. It is biased to the scholar's country of residence. We are working on adding more of them to our database.

4.3 Name Disambiguation

With the amount of data and data sources rapidly growing and expanding, it is essential for the large amounts of available data to be organized for analysis. Through the process known as Data Wrangling, unorganized and scattered data can be prepared for easy access and analysis. The datasets of google and goverment funding have to be cleaned because it contains many non-english characters and its messy datasets [15]. We use regular expression to remove the invalid special characters and translate phonetic characters to english alphabets. We designed and implemented Algorithms

10 to match the author names in Google Scholar with those in NSF/NIH/NASA data. This process helps to improve the quality of results.

4.3.1 Within the Google Scholar profile

A single Google Scholar profile might contain multiple variations of the authors name based on the middle name and initials. For example, consider the example Google Scholar profile of Ioannis Pavlidis. It contains four variations of his name in different publications.

- Ioannis T Pavlids
- IT Pavlids
- I Pavlids
- Ioannis Pavlids

We use the first initial and last name of an author to obtain the count of the number of publications in the panel.

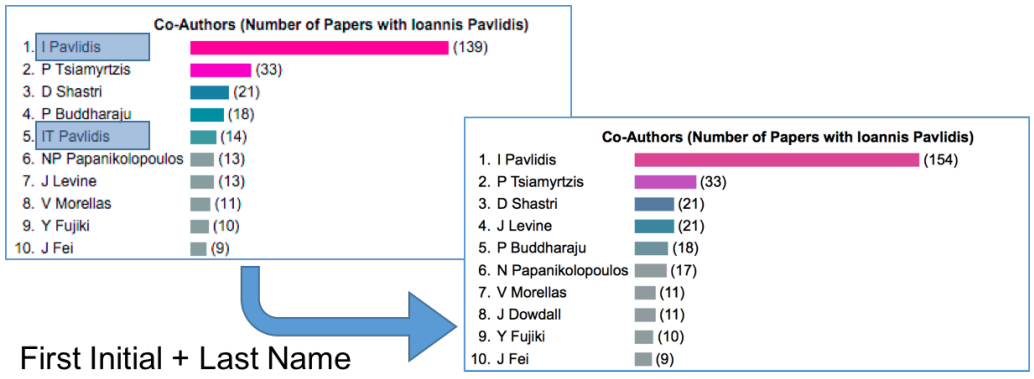

Figure 4.2: Example of how the name disambiguation algorithm works.

4.3.2 Between Google Scholar and Funding datasets

The funding datasets released from governments need to be cleaned because they are different data formats and structure. We cleaned the names by removing Sr., Jr., III, Ph.D., Dr., and so on. Then we need to match the names in the Google Scholar profile with those in the funding datasets. The algorithm is given in Algorithm 10. An example is visually depicted in Figure 4.3.

Algorithm 1 Matching the name between Google Scholar and funding datasets

procedure SEARCHING FOR AUTHOR NAME
    $googleFirstName \leftarrow$ first name in Google Scholar
    $googleLastName \leftarrow$ last name in Google Scholar
    $googleMiddleInitial \leftarrow$ middle initial in Google Scholar
    if $lastNameInFundingData = googleLastName$ **then**
        if $firstNameInFundingData = googleFirstName$ **then**
            if $googleMiddleInitial$ is null **then return** true
            else Search for $(middleInitial, googleFirstName)$ and $(googleFirstName, middleInitial)$
                if $found$ **then return** true
                else return false
    return false

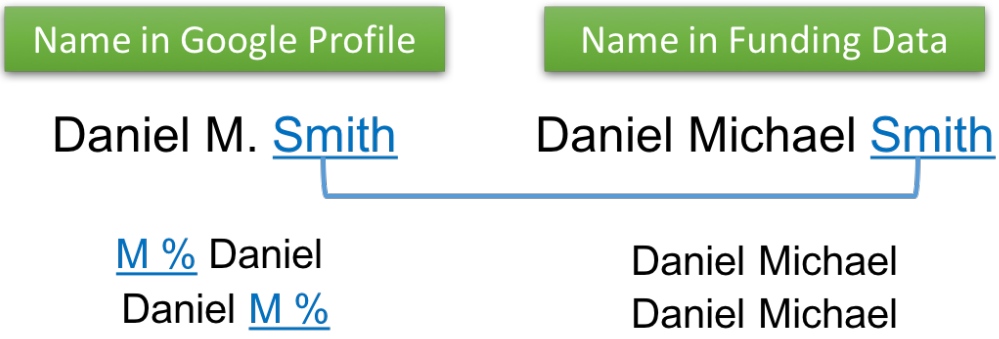

Figure 4.3: Example of matching the name in Google Profile with the name in funding data. Daniel M. Smith is considered as Daniel Michael Smith and Daniel Smith.

Chapter 5

Results

5.1 User Feedback - Usability Study

A total of 15 participants from various disciplines including Natural Sciences, Social Sciences, Life Sciences and Computer Science evaluated Scholar Plot. We asked each participant to review the interface and complete an online survey. Special care was taken to ensure that the participants had correct understanding about the visualization component before they began rating. The participants answered the questions on a Likert scale from 1 to 5 with 1 being strongly disagree and 5 being strongly agree.

Figure 5.1 illustrates the mean evaluation for each visualization component. Accuracy, usability, and understandability of Scholar Plot scored the highest ($\mu = 4.2$)

as it is very intuitive and can be used with minimal assistance. The highest positive feedback we received from many of the participants was the visual scheme of Scholar Plot. Another observation is that the participants agree to use Scholar Plot to evaluate themselves ($\mu = 4.1$). They suggested that Scholar Plot can be improved by adding more funding agencies. Overall, this evaluation indicated that Scholar Plot is a user-friendly tool that complements the CV which can be used to review a scholar's accomplishments. The survey has been approved by the University of Houston Institutional Review Board (IRB).

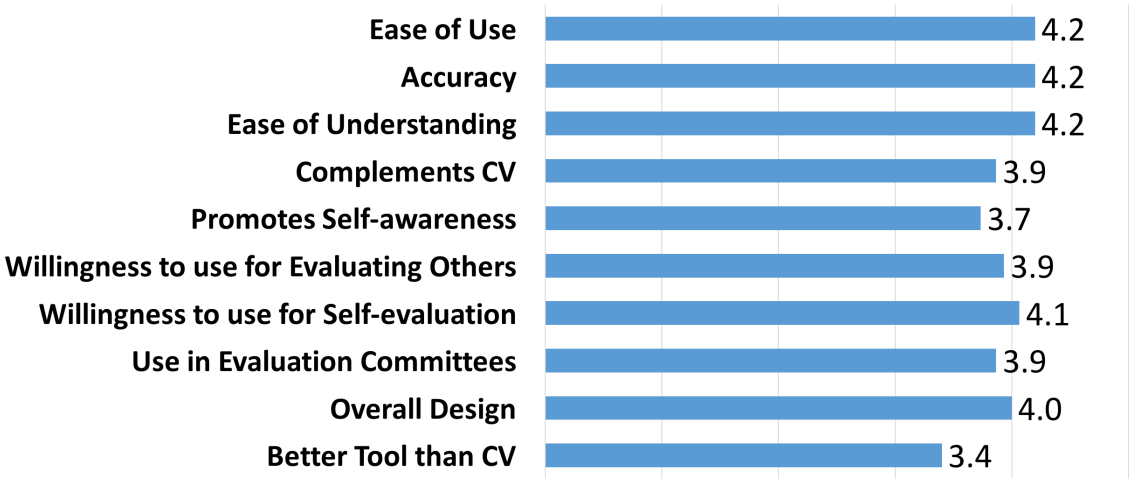

Figure 5.1: Mean evaluation of Scholar Plot. A total of $n = 15$ participants evaluated the survey.

5.2 User Feedback - Focus Group

We ran a focus group with 10 Principal Investigators and their post doctors at Northwestern University. The participant set included biologists, physicists, computer scientists, and social scientists. The focus group's suggestions are synopsized as follows:

Interface team science information. Participants wanted to see the number and intensity of collaborations for the depicted scholar.

Summarize highly cited papers. Participants wanted to see explicitly in a side panel the scholar's most popular papers.

Interface journal profile. Participants wanted to see the specific journals where the scholar publishes most often and their impact factors.

The participants believed that accessorizing the central publication graph with this additional information would support deeper instant comprehension without compromising the elegance of Scholar Plot's compact visual representation. Specifically, this additional interface would reveal the collaborative nature of the scholar's work, give hints if s/he is a regular in specific disciplinary journals or if publishes in a variety of journals (interdisciplinarity), and give the rank of these journals.

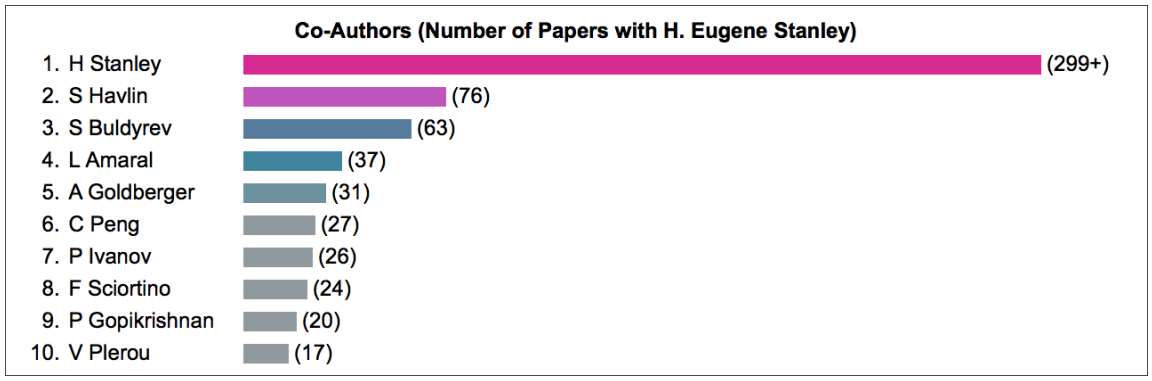

Figure 5.2: Panel listing the top collaborators with the selected scholar ranked by the count of the number of publications collaborated.

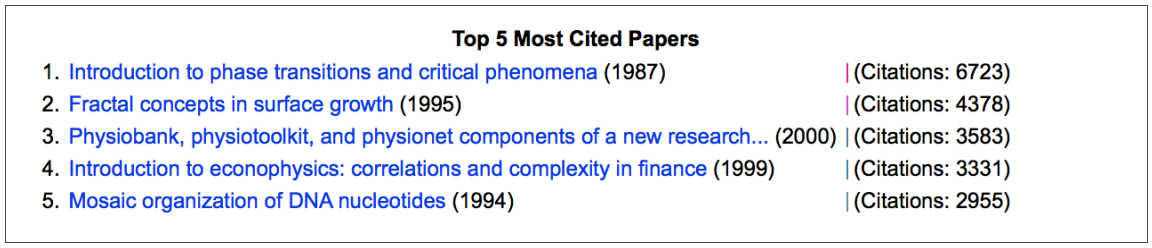

Figure 5.3: Panel highlighting the top 5 cited papers of the selected scholar.

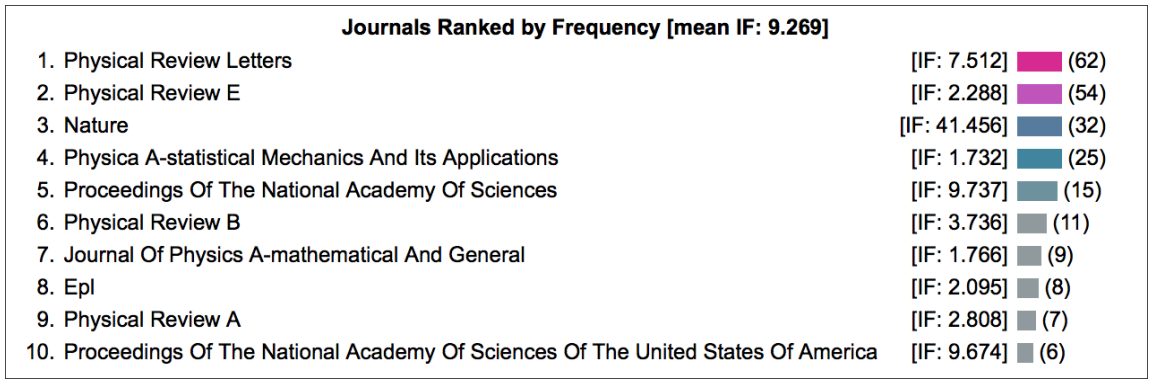

Figure 5.4: Panel displaying the top journals ranked by the frequency of publication.

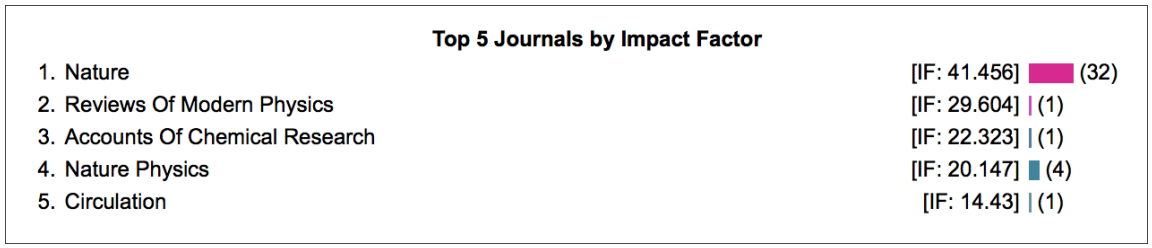

Figure 5.5: Panel showing the top 5 journals where the selected scholar published ranked by the impact factor.

5.3 Global and Local Bias Correction

Academic Garden reveals that some people stand out locally in low ranking departments such as the University of Houston (See Figure: 5.6) but are ordinary in the global scheme of visualization (See Figure: 5.7). The opposite is true for very high-ranking departments such as Massachusetts Institute of Technology (MIT) where,

because of a couple of outstanding people, others may appear unimportant locally (See Figure: 5.9) though they are quite good in their discipline (See Figure: 5.7).

Providing both visualization schemes to the user makes Academic Garden a useful tool for every academic department no matter how they are ranked. Rather than a log scale, which would unjustifiably elevate the lowest performing faculty and not adequate acknowledge the merit of the highest performing faculty, the local scheme displays a department using a linear scale. The local scale is dynamic and is adjusted to the maximum citation count of the department.

The global scale has two linear sections. The top section has a fixed minimum of 20,001 and a fixed maximum of 300,000, which is larger than the highest number of citations in the global dataset. The bottom section has a fixed range of 0-20,000 citations, which represents the 90^{th} percentile of the global dataset. The larger height of the bottom section displays the 90^{th} percentile of faculty vertically across Academic Garden instead of compressing it to the very bottom.

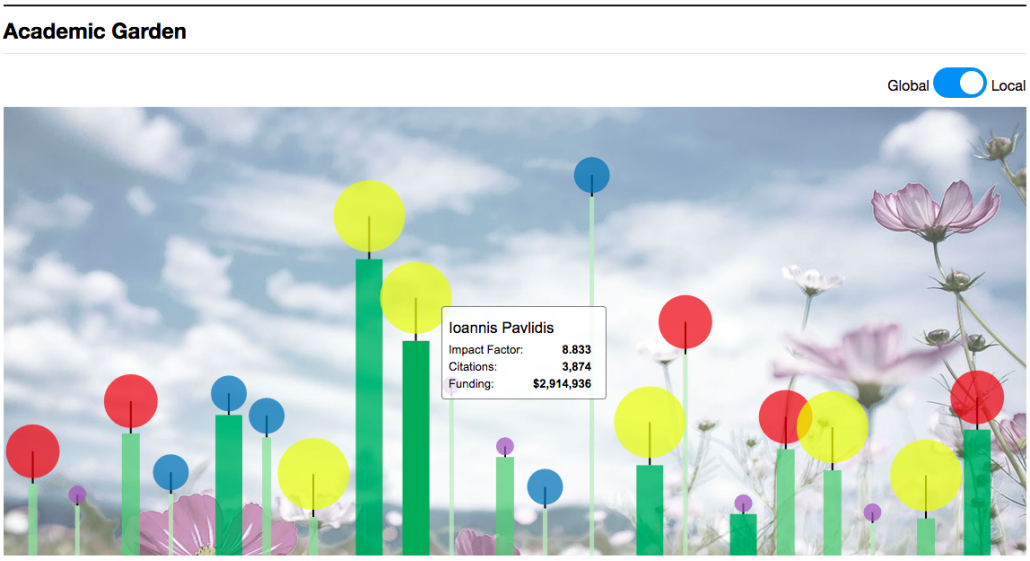

Figure 5.6: Local Scale: Department of Computer Science at the University of Houston

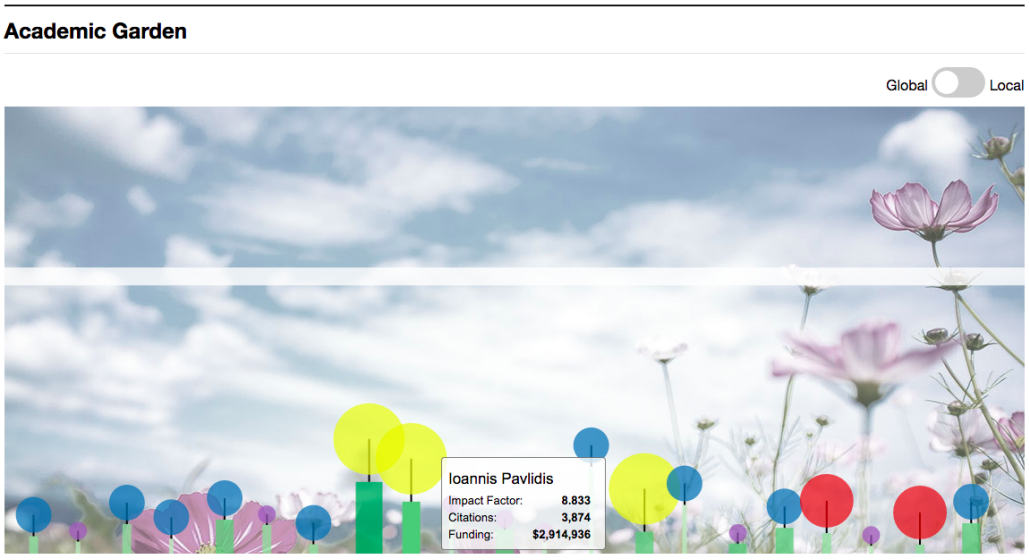

Figure 5.7: Global Scale: Department of Computer Science at the University of Houston

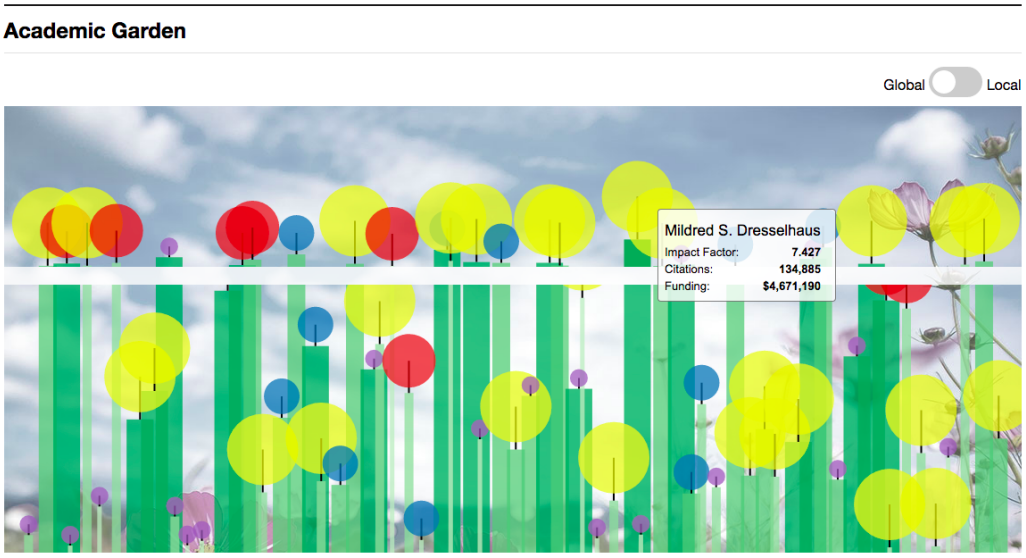

Figure 5.8: Global Scale: Department of Computer Science at the MIT

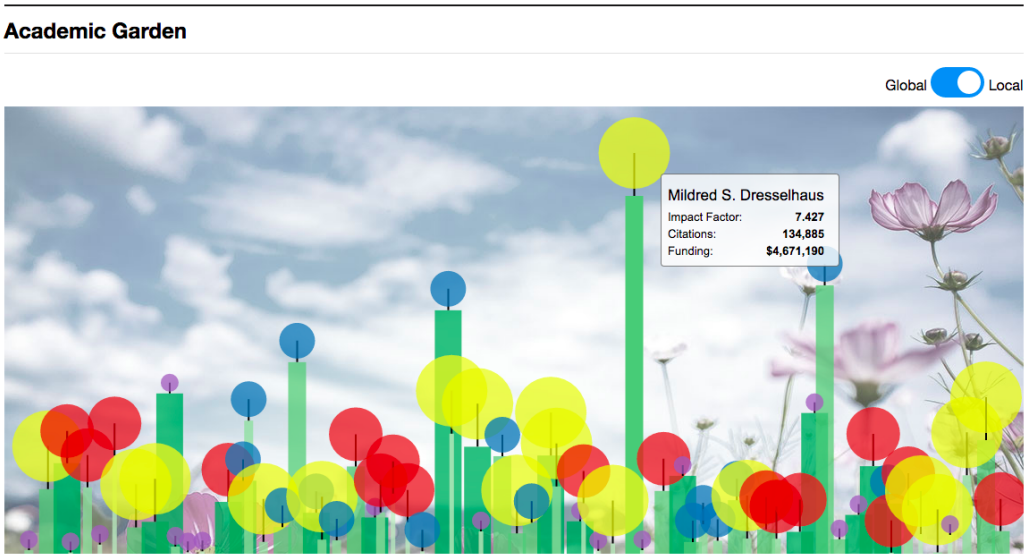

Figure 5.9: Local Scale: Department of Computer Science at the MIT

Table 5.1: The list of institutes in Computer Science by rank sourced from U.S. News [19].

Rank	Department of Computer Science
1	University of California, Berkeley
1	Carnegie Mellon University
1	Massachusetts Institute of Technology
1	Stanford University
5	University of Illinois at Urbana-Champaign
6	Cornell University
6	University of Washington
8	Princeton University
9	Georgia Institute of Technology
10	University of Texas, Austin

5.4 Data Analysis

For the validation of our design choices of Academic Garden, I used chaired faculty as the ground truth. Endowed Chaired faculty is considered a prestigious award in the United States. I ran linear models in R software [9] to understand the validity of the design with respect to Endowed Chairs.

The data has been collected in July 2016. This included total 14 different universities. The data consists of ($n = 248$) faculty from Computer Science and ($n = 152$) from Biology from the top 10 schools according the US News Report 2015 [19]. The data of chaired faculty consists of ($n = 61$) chaired professors from Computer Science and ($n = 32$) from Biology in top 10 schools.

The three criteria we used in the Academic Garden are citations, impact factor, and funding. We computed the quartiles for these 3 criteria based on the local department faculty, as well as the global scale considering all the faculty from the

Table 5.2: The list of institutes in Biology by rank sourced from U.S. News [19].

Rank	Department of Biology
1	Harvard University
1	Massachusetts Institute of Technology
1	Stanford University
4	University of California, Berkeley
5	California Institute of Technology
5	Johns Hopkins University
7	University of California San Francisco
7	Yale University
9	Princeton University
10	Cornell University

same discipline. We obtained the discipline information from the Classification of Instructional Programs (CIP) codes from The National Center for Education Statistics designed the Classification of Instructional Program [7].

For each faculty, we computed the quartile to which he belongs to for each of the three criteria in the local and global scales. We also computed a variable to determine if a faculty belongs to either one of the top 3 criteria. For Computer Science, this variable significantly predicts chaired faculty ($p < 0.05$) i.e, we can predict a faculty is chaired if he belongs to the top quartile locally in either of the three criteria.

The values are shown in Figures 5.10. In this model, quartiles are calculated with respect to the department which the faculty belongs to.

All three criteria can be considered as separate factors. In computer science, citations are highly significant ($p < 0.001$) while the mean impact factor is not significant. This is because computer science faculty does not publish so much in

```
summary(glm(is_chair~local_top_q_any_of_three, family = binomial))
```

```
##
## Call:
## glm(formula = is_chair ~ local_top_q_any_of_three, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8567  -0.8567  -0.6039  -0.6039   1.8930
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -1.6094     0.2582  -6.233 4.57e-10 ***
## local_top_q_any_of_three   0.7959     0.3166   2.514   0.0119 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 276.70  on 247  degrees of freedom
## Residual deviance: 270.02  on 246  degrees of freedom
## AIC: 274.02
##
## Number of Fisher Scoring iterations: 4
```

Figure 5.10: Screenshot of a result of Linear Model in R

journals. The funding is also not significant because our funding sources (NSF/NIH/NASA) do not include most funding sources which computer science faculty get funding from, for example, DoD (United States Department of Defense) and DHS (United States Department of Homeland Security).

However, in the case of Biology, the funding quartile is significant ($p < 0.01$) because of the funding dataset includes NSF, NIH from where most the Biology grants are from (Figures: 5.11). Also, the Impact Factor quartile is significant ($p < 0.05$) because they publish more in journals.

According to the results of linear model, the data analysis validates the design choice of the three criteria for the visualization, and it is exactly mirroring visualization with quartiles values.

```
summary(glm(is_chair~local_q_t_funding, family = binomial))
```

```
##
## Call:
## glm(formula = is_chair ~ local_q_t_funding, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9142  -0.7435  -0.4758  -0.4758   2.1140
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.1680     0.4366  -0.385  0.70037
## local_q_t_funding  -0.4883     0.1782  -2.740  0.00614 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 156.45  on 151  degrees of freedom
## Residual deviance: 148.32  on 150  degrees of freedom
## AIC: 152.32
##
## Number of Fisher Scoring iterations: 4
```

Figure 5.11: Screenshot of a result of Linear Model in R

Chapter 6

Conclusion

I described a visualization method that complements the information contained in a researcher's Google Scholar page and summarized by her/his h-index. I provided a visualization scheme which summarizes the current measures. It introduces a bias due to funding agencies in the United States. One can draw deeper conclusions that are not supported by the h-index alone and cannot be derived from the Curriculum Vitae or the Google Scholar page, unless a significant investigative effort is undertaken. The qualitative panels, statistical values like mean impact factor, temporal plots, and the tooltips provide such useful insights. Our user study also supports this.

Scholar Plot works at three levels - the individual, the department, and the college. The individual (base) level captures in a figure three key indicators of academic prowess: citation impact, the prestige of publication venues, and research funding. These indicators scale up in the department & college (aggregate) levels of Scholar Plot as pie charts, revealing at a glance the relative contributions of entities from

the lower echelon.

The basic idea behind Scholar Plot is to facilitate an instant deeper comprehension regarding different strengths of academic records, supporting the work of evaluation committees, and the curious academic in search of an advisor or department. One of Scholar Plot's strengths is that it draws data from open sources that are inclusive. However, it is a technical problem because Google Scholar - a key open source used by Scholar Plot - does not offer an application programming interface (API). For the base level of Scholar Plot, we solved this problem with sophisticated data scraping assisted by a simple one-time wiki function: if the individual sought by the user is not recognized by Scholar Plot, Scholar Plot asks the user to copy and paste the targeted individual's Google Scholar URL. Scholar Plot will remember it thereafter by automatically scraping the scholar's data every time a user requests it by name. For the department and college levels, a wiki function is also available to request the information of the departments at `https://goo.gl/RHsuJu`.

Not only that, I described Academic Garden (AG), which is about individual academics, departments, colleges, and any other academic group visualization. Academic Garden uses the flower metaphor to visually articulate performance for academic entities. The width of the flower's stem is commensurate to the academic funding the entity received ('juice conduit'). The height of the flower's stem is commensurate to the impact of the entity's intellectual products ('visibility'). The diameter of the flower's disc is commensurate to the prestige of the venues where these products appeared ('fancy factor').

For the validation of the design choices of Academic Garden, I used Endowed

Chaired faculty as the ground truth. Endowed Chaired faculty is considered as a prestigious award in the United States. The data analysis using faculty from the Computer Science and Biology departments of the top 10 schools in the United States indicates that chaired faculty can be predicted using the three merit criteria of citations, impact factor, and funding. Our scheme is exactly mirroring the visualization with quartiles values.

The Scholar Plot and Academic Garden are likely to have a broad appeal because it is useful for evaluating committees, and it is available online for free at `http://www.scholarplot.com`.

Chapter 7

Appendix

7.1 Usage of Scholar Plot

In this section, I will explain how to access and use Scholar Plot. This includes searching for a scholar from Google Scholar, inserting a scholar into our system, obtaining results, and the scholar's profile URL.

7.1.1 Searching for a scholar

To visualize the accomplishments of a scholar, type the name of a scholar in the search box. As you type, Scholar Plot will attempt to match your query to the names of scholars in our system.

If the result of a search produces no results, Scholar Plot will prompt you to enter the URL of the person's Google Scholar Citations Profile. Instructions for finding a

Google Scholar URL can be found.

7.1.2 If a name cannot be found

When a name cannot be found in our system, Scholar Plot will prompt the user to enter the URL of the person's Google Scholar Citations Profile. This URL can be found using the search bar on the Google Scholar website (here).

If the names of more than one scholar match the query, you will need to locate the correct scholar in the search results.

7.1.3 Google Scholar Author Search Results

From the person's Google Scholar Citations Profile page, copy the URL from your web browser's address bar.

7.1.4 Obtaining the Google Scholar Profile URL

Return to Scholar Plot and click the 'Submit' button. The scholar's information will then appear and their name can be used in future searches on Scholar Plot and Scholar Compare.

Figure 7.1: Usage of Scholar Plot - Type the name of a scholar in the search box.

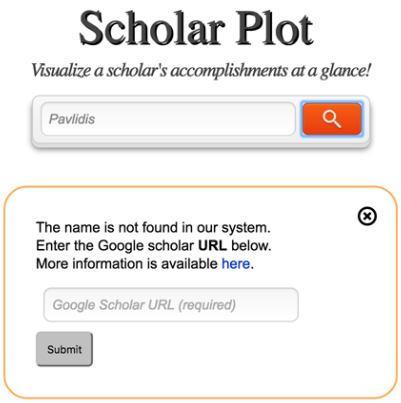

Figure 7.2: Usage of Scholar Plot - No Results in Scholar Plot Search in Scholar Plot System.

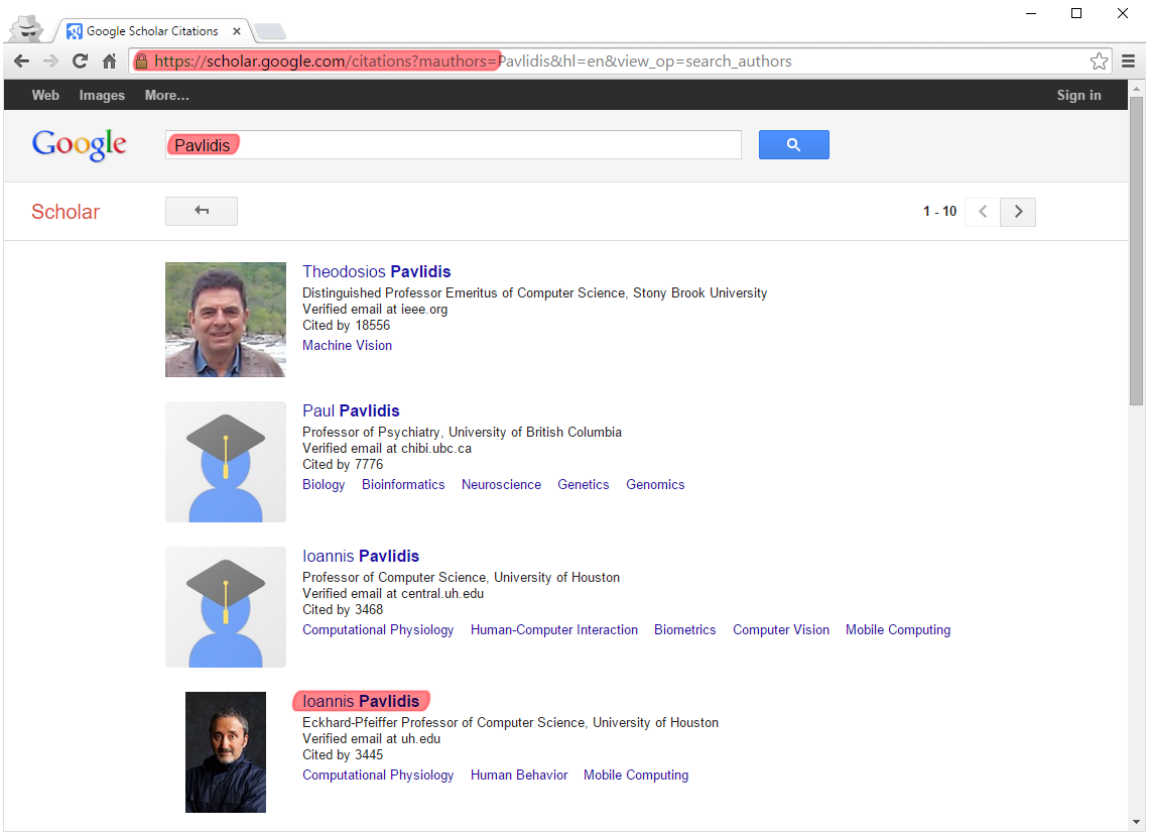

Figure 7.3: Usage of Scholar Plot - Searching a scholar profile in Google Scholar.

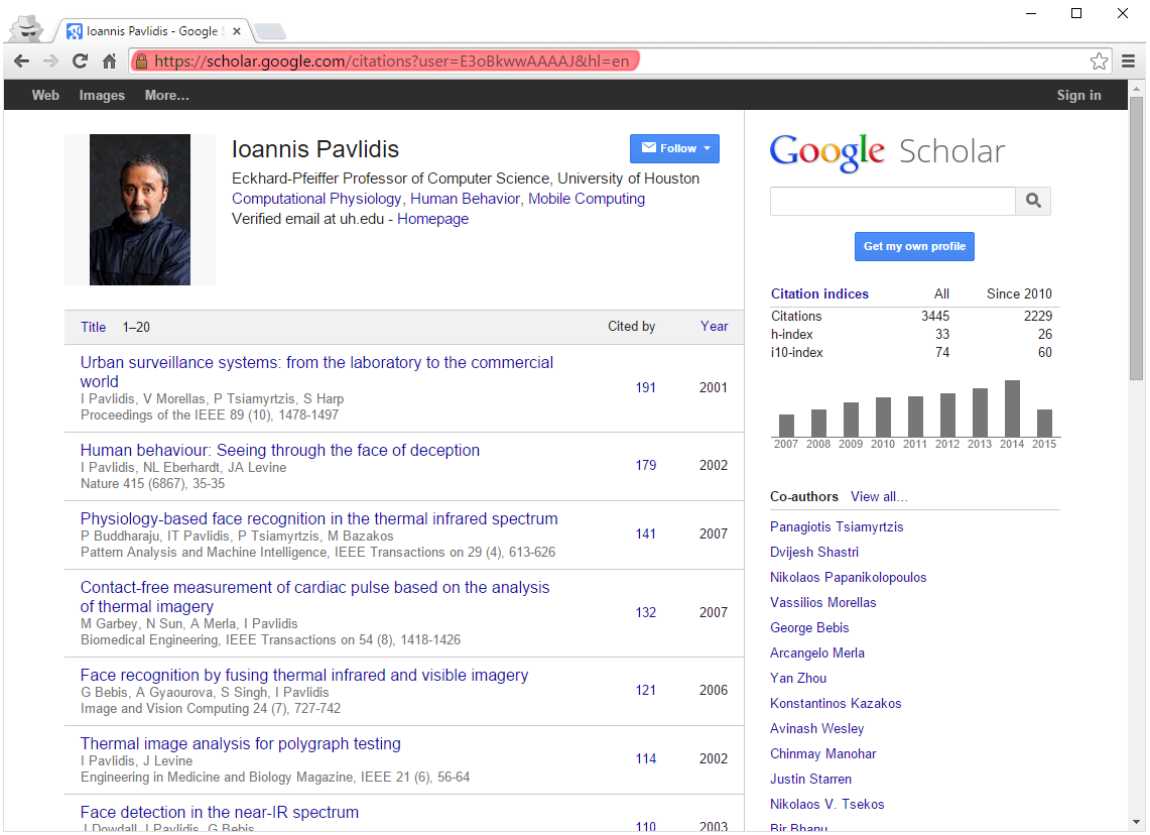

Figure 7.4: Usage of Scholar Plot - Copying the Google Scholar Citations Profile URL.

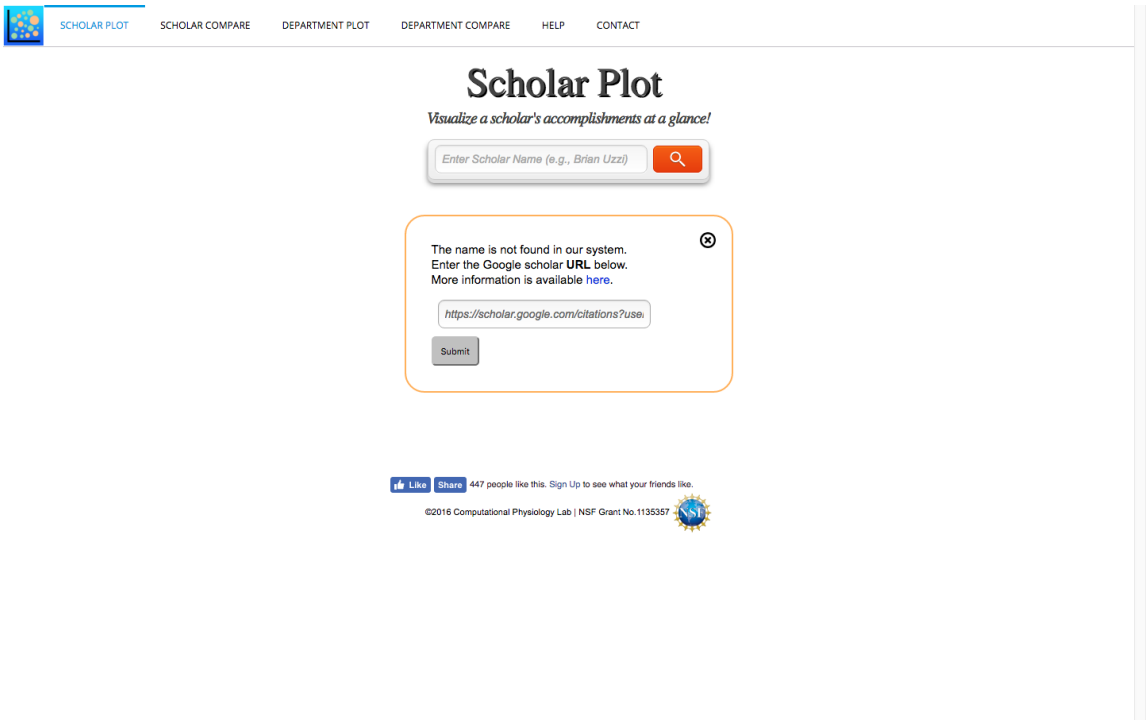

Figure 7.5: Usage of Scholar Plot - Pasting the Google Scholar Citations Profile URL.

Bibliography

[1] N. Aeronautics and S. Administration. Research.gov - research spending & results. https://www.research.gov/research-portal/appmanager/base/desktop?_nfpb=true&_eventName=viewQuickSearchFormEvent_so_rsr, 2016. [Accessed on 2016].

[2] L. Bornmann and H.-D. Daniel. The state of h index research. is the h index the ideal way to measure research performance? *EMBO reports*, 10(1):2–6, 2008.

[3] M. Bostock. D3: Data-driven documents. d3 (or d3.js) is a javascript library for visualizing data using web standards. https://d3js.org/, 2016. [Accessed on 2016].

[4] K. Brner. *Atlas of Science: Visualizing What We Know*. The MIT Press, 2010.

[5] C. Chen, R. J. Paul, and B. O'Keefe. Fitting the jigsaw of citation: Information visualization in domain analysis. *Journal of the American Society for Information Science and Technology*, 52(4):315–330, 2001.

[6] Esevier. Scopus - the largest abstract and citation database of peer-reviewed literature. https://www.scopus.com/. [Accessed 2016].

[7] T. N. C. for Education Statistics designed the Classification of Instructional Program. Classification of instructional programs (cip). https://nces.ed.gov/ipeds/cipcode/Default.aspx?y=55. [Accessed on 2016].

[8] N. S. Foundation. Nsf award search: Download awards by year. http://www.nsf.gov/awardsearch/download.jsp, 2016. [Accessed on 2016].

[9] T. R. Foundation. R: The r project for statistical computing. https://www.r-project.org/. [Accessed on 2016].

[10] E. Garfield. The history and meaning of the journal impact factor. *Jama*, 295(1):90–93, 2006.

[11] Google. Google scholar. https://scholar.google.com/. [Accessed on 2016].

[12] M. Grushinskiy. Xmlstarlet command line xml toolkit. http://xmlstar.sourceforge.net, 2016. [Accessed on 2016].

[13] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–16572, 2005.

[14] J. John. hxseven/htmlsql: htmlsql is a experimental php library which allows you to access html values by an sql like syntax. https://github.com/hxseven/htmlSQL, 2016. [Accessed on 2016].

[15] H. Kang, L. Getoor, B. Shneiderman, M. Bilgic, and L. Licamele. Interactive entity resolution in relational data: A visual analytic tool and its evaluation. *IEEE transactions on visualization and computer graphics*, 14(5):999–1014, 2008.

[16] J. Kaur, M. JafariAsbagh, F. Radicchi, and F. Menczer. Scholarometer: A system for crowdsourcing scholarly impact metrics. In *Proceedings of the 2014 ACM Conference on Web Science (WebSci 2014)*, pages 285–286, Bloomington, Indiana, June 23-26, 2014. ACM.

[17] L. Leydesdorff. Visualization of the citation impact environments of scientific journals: An online mapping exercise. *Journal of the American Society for Information Science and Technology*, 58(1):25–38, 2007.

[18] P. M. McDonough, A. Lising, A. M. Walpole, and L. X. Perez. College rankings: democratized college knowledge for whom? *Research in Higher Education*, 39(5):513–537, 1998.

[19] U. News. Best colleges — college rankings — us news education - us news. http://colleges.usnews.rankingsandreviews.com/best-colleges, 2016. [Accessed on 2016].

[20] N. I. of Health. Exporter data catalog. http://exporter.nih.gov/ExPORTER_Catalog.aspx, 2016. [Accessed on 2016].

[21] N. Partners. Nvd3: A reusable charting library written in d3.js. http://www.nvd3.org, 2016. [Accessed on 2016].

[22] T. Reuters. Thomson reuters - ip & science - web of science. http://ipscience.thomsonreuters.com/product/web-of-science/?utm_source=false&utm_medium=false&utm_campaign=false. [Accessed on 2016].

[23] A. Robecke, R. Pryss, and M. Reichert. Dbischolar: An iphone application for performing citation analyses. In *CAiSE Forum-2011*, number Vol-73 in Proceedings of the CAiSE'11 Forum at the 23rd International Conference on Advanced Information Systems Engineering. CEUR Workshop Proceedings, June 2011.

[24] D. Toad. Toad for oracle — oracle database tools — sql development & administration — dell software. `http://software.dell.com/products/toad-for-oracle/`. [Accessed on 2016].

[25] E. Vardell, T. Feddern-Bekcan, and M. Moore. SciVal experts: A collaborative tool. *Medical Reference Services Quarterly*, 30(3):283–294, 2011.

[26] J. S. Yi, Y.-a. Kang, J. T. Stasko, and J. A. Jacko. Understanding and characterizing insights: How do people gain insights using information visualization? In *Proceedings of the 2008 Workshop on BEyond Time and Errors: Novel evaLuation Methods for Information Visualization*, BELIV '08, pages 4:1–4:6, New York, NY, USA, 2008. ACM.