Interactive Qualifying Project

submitted to the faculty of

Worcester Polytechnic Institute

in partial fulfillment of the requirements for the degree of bachelor of science

A History of Physics at WPI

Christopher M. Pierce

Supervised by
Germano Iannacchione

April 18, 2017

Abstract

Two novel datasets on historical faculty publications and career trajectories at Worcester Polytechnic Institute are produced. These datasets are then studied for trends and patterns within the academic history of the institution. From these investigations the context of WPI in the greater physics community is explored.

Acknowledgments

The author would like to thank the WPI library and especially Michael Kemezis from the archive for providing access to the original materials used to generate the datasets in this paper. Without their support this project would not be possible.

Contents

List of Figures

List of Tables

1 Introduction

As best put by George Santayana, "Those who fail to learn from history are doomed to repeat it." Worcester Polytechnic Institute (WPI) has been involved in the study of physics since its founding in 1865 and as of yet no real history has been compiled. This means that current students, researchers and faculty have no way of understanding the failures and successes of their predecessors outside of the occasional verbal history from one of the longer tenured members. It is then difficult or impossible to learn from the past or, equally importantly, understand the context of current events in relation to the past. A history of physics at WPI will not only enable a better understanding of WPI's academic legacy, but will provide lessons for the future of research performed at the Institute.

Our story begins with the founding of the university by John Boynton and Ichabod Washburn. While the two independently wished to create an institution of higher learning, their visions differed fairly drastically. Boynton sought to "elevate the position of the farmer, the mechanic, and the manufacturer, not necessarily teach him how to be one."[1] On the other hand, Washburn was looking to create a school to train the next generation of skilled workers through an apprenticeship approach. It was Seth Sweetser, a local pastor, who by coincidence was approached by both individuals for his advice.[2] Sweetser drafted a letter to 30 Worcester businessmen on behalf of the two to secure funding for a "Free School for Industrial Science."

In 1865, the university was created legally.[1] It quickly became known for producing good quality engineers of all types.[2] In the late 1960's a new type of undergraduate curriculum was created known as the WPI plan. Focusing on three projects to be completed as a student, it marked a serious deviation from more traditional engineering programs. The class of 1972 became the first to contain women, as WPI was founded as an all male college. The remaining portion of the 20th century saw WPI grow to its current position in the world.

In the time since WPI's founding, the study of physics underwent paradigm shifting changes. Einstein published his famous works on special relativity in 1905 followed by his theory of general relativity in 1916. At the same time quantum mechanics was in its infancy with the Schrodinger and Heisenberg formulations being created in the mid 20's. Relativistic quantum mechanics followed quickly in 1928 with work by Paul Dirac and developed into quantum electrodynamics in 1948. These fundamental discoveries led to the creation of entire subfields in their wake bringing us to the present state of affairs.[3]

2 Methodology

The history of physics at WPI was studied by restricting the project's focus to two tangible subjects: the faculty and the research they performed. Although there were many potential subjects that could have been studied, these were chosen for a couple of good reasons. The first is that these subjects conform to the typical person's idea of what the history of an academic institution such as WPI entails. By beginning here, one can hope to uncover all aspects of WPI's history considered to be important by traditional standards. Another reason for studying these two subjects was the availability of reliable material related to them. Without high quality data, results could not be confidently gathered. Finally, out of the various subjects considered these two appeared the most likely to generate new insights.

Knowledge of the various characters and personalities that made up the WPI physics department over the years forms an important foundation to any understanding of its history. The goals in studying WPI's faculty are to understand the culture of the department and how it has changed over time. Research and work are always intertwined with the personal interactions of those performing it. It is the hope that by understanding these interactions and the ways that they have changed, a greater understanding of the past may be achieved. This may then be directly applied to achieve a better understanding of WPI's academic history.

The products of research performed at WPI are equally important to the discussion of WPI's history as the people at the institute. The goal in studying this subject is twofold. First off, it should uncover the various trends and patterns in how research has been performed. This is important to the goals of the project as it will have the opportunity to reveal lessons on how to encourage good quality research. The second goal in studying this subject is to determine the major historical events in research at WPI. This will help uncover WPI's legacy as it applies to the field of physics.

2.1 Enumerating WPI Faculty

The first challenge in studying the people of the WPI physics department was finding reliable sources of data on them. The first attempts at investigation were by contacting various groups that maintain records at WPI. This included the physics department itself, WPI's office of the provost and WPI's alumni relations office. The physics department unfortunately did not keep significant records of past faculty and those they did keep only spanned the period of a few decades. The provost's office had the potential to offer more information, however the information they could find was primarily focused on tax information and was out of the scope of this project. Finally, although the alumni relations office did try to keep records of past members of the community, their resources on faculty members were extremely limited and did not provide information of use to this project.

This lack of information on WPI faculty members meant that other options had to be explored. Further inquiry led to the discovery that the WPI archives held copies of all course catalogues published at the school. Between their covers were listings of information on WPI faculty members enabling research to progress. Entries on faculty from the course catalogues typically contained the following information:

Name The name of the faculty member. The format of the name changed by a significant amount depending on the year. In certain years, the name would be shortened to ”¡first initial¿ ¡last name¿” while in others the name would be printed in its entirety. The format also appeared to depend on individual preference, especially with regards to middle names. Later analysis would require that a single faculty member maintained the same name through their tenure at WPI. Fortunately there were no instances of names being changed, for instance because of marriage. Therefore, a policy was adopted to take the name with the most information available. That is the name with the most sections and in their fully expanded form.

Qualifications For most faculty members this section listed the degree that they held at the time of the catalogue's publication. However, certain faculty members in the past were qualified for work in the department by other means. As an example, one member was listed as having passed his Staatsexamen which is a German government run licensing exam for professionals and is similar to a Master's degree. Certain catalogues included a list of all qualifications up to the date of the printing of the catalogue whereas others only included the highest qualification. For the purpose of this article, it was of interest to find the highest degree held in a given year.

Notes There were certain events or pieces of information that were included in a few catalogues that did not fit into any other section. This mostly included footnotes and addenda. Described in this section were leaves of absence, special faculty appointments and a few cases where a faculty member died and was still listed in the catalogue.

Title The title held by the faculty member at the time of the catalogue's publication. It is interesting to note that there does not appear to have been a standard in place for faculty titles until the 20th century. Early titles also changed from year to year and it appeared to be common for faculty to move around between departments. This behaviour stopped in the early 20th century and a common path for promotion appeared to form.

Address of residence The early course catalogues also included a section listing the permanent residences of the faculty. This information disappeared in the early 20th century and was not collected as it was out of the scope of the project and would have taken a large amount of time away from other more significant sources of information. One interesting change that was noted, although the data was not kept, is that many of the early professors lived as room mates or in very close proximity to one another. This likely changed due to the national migration of people away from cities in the mid 20th century.

Departmental association Finally, the catalogues also included information about departmental associations. This was not very important to the goals of this project as we were focused specifically on the physics department and not the overall makeup of faculty at WPI. In the early years of the university, the course catalogues did not separate faculty members into different departments. In order to tell what the individual did it was necessary to look at their title. In the mid 20th century there was a transition from listing the faculty as a whole to listing the faculty as they are split into departments.

As indicated by the previous discussion, the course catalogues evolved significantly over the years. Many pieces of information available in certain years would not be available in others or the formatting of information would change from year to year. Another example not demonstrated before is that graduate students would appear in certain course catalogs and then not be listed in subsequent years. These difficulties were overcome by simply gathering all available information at this point in the project and developing methods of analysis that would handle the changes later. One other interesting challenge was that the scheme used to label course catalogues actually changed part way through the 20th century from using the calendar year of the start of the academic year to the calendar year that the academic year ended. In order to deal with this, the entries were denoted internally using both the ending calendar year and the starting calendar year.

Each entry was transcribed by the author into a spreadsheet program. The information transcribed was: the year of the entry, the name of the faculty member, their highest qualification, their title, and any notes in the catalogue that year. By the end of this phase of the project, a rather unwieldy document was generated with the information contained in all of the course catalogues. To enable future analysis, it was decided to transfer the information into a database management system (DBMS). The DBMS chosen was MariaDB, a common open source application well suited for the task at hand. Using a program written in the python programming language, the document was transferred into a table in the database.

2.2 Collecting Publication History

As with the investigation of faculty, the first resources to be investigated were those at WPI. It was the hope that local sources could help provide selectivity and filter out much of the data not useful to this study. The WPI library and physics department were both asked for information about publications made by faculty at the university. Discussion with library staff members indicated that there were some collections of this kind of data available. However, these were not very complete and did not lend themselves to analysis using modern techniques. In order to use them, they would need to be digitized, a process that would take an extraordinary amount of time given the volume of data. The physics department kept no records of publications at the institute.

The next step was to look for external collections of publications. Fortunately for this project, journal publishers keep good quality records of past publications and in recent years there has been a major effort to make past publications accessible digitally. It was discovered that by far the largest publisher of research performed in physics is the American Physical Society (APS). This organization was was founded in 1899 with the purpose "to advance and diffuse the knowledge of physics." At the time of this writing they have over 50,000 members and, of more interest to this article, 13 peer-reviewed research journals spanning the whole of the subfields within physics. Another positive attribute of this publisher is that they have invested large amounts of time and resources into creating digital archives of their work. This means that much of the effort required for the completion of this project was already finished.

The first challenge in retrieving as much information on the publications as possible was determining how to filter out all of the unrelated information. The APS database included many more papers than just those published by faculty members at WPI. In fact even for those published by those who were affiliated with WPI at one point in their career, they may not have have authored the paper while they were on staff. The two major criteria that were used to select potentially useful articles were the author list and the date of publication. In order to avoid losing data in the collection process the guidelines used for pulling information from the APS database were kept as loose as possible.

Due to the previous work in the project on collecting information about faculty members in the WPI physics department, a list of potential author names had already been assembled. One difficulty encountered when trying to collect all of the articles from WPI related authors was that the name and even the format of the name of the author may not be exactly the name in the list that was being used. The APS query tool could recognize when initials were to correspond to a full name, but could not go the other way. I.E. a search for "J. Doe" may turn up an article written by John Doe whereas a search for "John Doe" may miss an article written by "J. Doe." For this reason the following search procedure was adopted:

1. Shorten the name in question to the form "[first initial] [last name]"

2. Query all papers with the shortened name on the APS database

3. For each returned publication:

    (a) Retrieve the list of authors

(b) Compare with the full author's name from the query with the names in the author list

For example if the APS database was to be searched for an author named John Doe, the name would first be shortened to the form ”J. Doe.” All articles matching this name would then be retrieved from the APS database including some which may not actually be written by John. The author list of each article is then compared with the originally specified name, in our case ”John Doe.” As an example, the query may have returned an article with the sole author ”Jane Doe.” Comparison with the original name allows this article to be eliminated from the search results while still capturing articles written by ”J. Doe” that would not be returned in searches for ”John Doe.” Further selection was performed during analysis by ensuring that the date of publication for works of research matched the period during which the potential author was affiliated with WPI.

Now that articles could be filtered based on their association with authors at WPI, the next step was to begin collecting the data in a form that would enable later analysis. For each query, the APS database search tool returned in a human readable format, a set of URLs to webpages on the articles matching the query. The webpages then contained a large amount of useful information about the articles, the authors, and the publication. As with the query results, these pages were returned in a human readable format which did not easily lend itself to the collection of data using an automated system. The information returned on articles included the following:

Title The title of the publication as it was printed by the journal

Publication Date The date the publication was published by the APS

Link to PDF of Article A URL pointing to a PDF copy of the publication. The project was interested in the metadata surrounding publications and not their contents. Therefore while the link was recorded, no PDFs were downloaded.

Citation Count The approximate citation count for the publication at the time the query was performed. This may not have been accurate, especially for older articles as it depends on the availability of digital copies of materials that refer to the publication.

DOI Digital Object Identifier. A unique way to locate the publication used in many organizational systems and defined by ISO.

Journal A unique shortened version of the name of the journal in which the publication appeared. For example, the name ”PR” was used to refer to the journal ”Physical Review.”

Volume The volume of the journal in which the publication appeared.

Page Range The range of pages inside the journal where the publication can be located.

Publication Type The type of publication, whether it was an article, a letter to the editor, etc.

Article ID An identifier used to refer to the publication internally by the APS.

APS Score An unidentified score associated with each publication by the APS database. It is believed that the score is simply a ranking of how closely each publication matched the query that located them for sorting purposes.

Author List The list of authors appearing on the publication.

References List A digital copy of all of the references associated with the publication. This is useful to the APS as it enables them to compute citation counts and rank their publications in that manner.

The next step in the process was collecting the information in a format usable during the analysis stage of the project. Like the collection of faculty information from the course catalogues it would have become extremely unwieldy to use something like a spreadsheet program to store the information. It was decided that a database would be created as it is the standard method by which large amounts of information are stored and quickly retrieved at the time of this writing. As with the collection of information from the course catalogues, the database management system MariaDB was selected due to its wide acceptance and the fact that it is open source. A piece of software was written in the programming language Python to quickly collect information from each query generated by the professors name. The overarching steps in the software package were:

1. Create a list of unique faculty names from the list of all course catalogue entries
2. For each faculty name do the following:
    (a) Query the APS database using the name shortening method described above
    (b) Parse the list of URLs pointing to articles on the returned webpage
    (c) For each returned URL, do the following:
        i. Parse the page for each field described above
        ii. If the selection criteria are met, insert the fields into a "article" table in the database
    (d) Pause for a few seconds
    (e) If there are more pages in the query, continue to the next
3. Log the query as being successfully completed in a "query" table in the database

Several challenges were posed while writing the software to collect information from the APS database. The first was being able to accurately parse the human readable content into a form usable by a computer and that may be analysed later. Fortunately, much of the data was included on the webpage in a format called JavaScript Object Notation (JSON). JSON is a commonly used method for passing information between technologies over the internet and JSON parsing is already implemented in a robust fashion in python. It was simply necessary to locate sections of the webpage that identified where JSON data was located. The strings could then be pulled out and fed into the existing JSON parser which would return the required data to be inserted into the "article" table.

The other challenge was that WPI has existed for a long amount of time which means that it has had many different faculty members over the years. The number of queries required to investigate every faculty member associated with the university is therefore quite high. It was necessary to add features to improve the reliability of the program. That way, in the event that the program was terminated abruptly or an error was made, the program could begin where it left off or the information associated with the offending queries could be removed. The solution was to include a table in the database with logs of all of the queries performed. Then, each entry into the "article" table had an identifier pointing to the query that produced it. A database maintenance program was developed that would clean up information that did not have a valid query associated with it. That way, a query could be removed, and all of the data associated with it cleaned up. The program could also check the log before executing a query to see if the search has been performed before and not continue if it has.

One final note on the software is that after every query of a page, the program halted for a few seconds. The reason for this is that webservers are typically configured for an average human user. That is someone who will perform a search and read the page for a few seconds to a minute before performing a new search. Running a program that performs many searches each second would put unnecessary stress on the APS servers. For this reason, the rate at which queries were performed was limited to almost the point of a human performing searches.

2.3 Analysis

The first difficulty encountered while trying to perform an analysis on the information was the sheer size of the dataset. If one imagines the number of faculty that have ever been associated with the WPI physics department and the average number of papers written by each individual, WPI's total publication count is quite high. Therefore, the size of the metadata associated with those publications is large and queries on that data are expensive in terms of computing power. Fortunately, the decision to use a database to store the information paid off greatly as database management systems are designed to handle such requests efficiently. The database management system also interfaced well with a number of programming languages that would enable the creation of data relations.

As before, the programming language python was selected to interface with the database management system MariaDB. The first path to analysing the data was to look at time series data. For the department/faculty related information this meant looking at trends over the years indicated by the course catalogues and for publications this meant looking at changes with respect to the indicated date of publication. For each year the department has existed, the database was queried for the quantity of interest in that year. These numbers were stored in memory as a list and associated with their year. Once all years had been queried, the list was output to a file in a format compatible with the program Gnuplot.

Gnuplot is an open source piece of software designed for plotting data such as that which is being used in this project. Plots are one of the simplest ways to locate trends and patterns in data and, at the time of this writing, humans are better than computers at spotting them in most types of data. The types of plots used in the project were:

Line Plot A plot charting the location of each point in the Cartesian plane with lines connecting subsequent points.

Pecentage Area Plot A plot where each vertical cross section is split into a number of colors and the length of the colored section relative to the total vertical axis represents a percentage.

Histogram A graphical representation of the relative frequency of a number of discrete objects.

The time series plots were created by considering the set of database entries associated with the given year as a statistical distribution. Then, various quantities such as the mean, variance, and number of data points were computed for each distribution and output as time series data. Taking the mean of the distributions generated information on how the average number of citations garnered on each paper changed over the years and the total number of data points created plots such as the total number of publications in a given year. Variance was mostly used for determining the reliability of trends as it is can be interpreted as how scattered the data is.

Histograms were then formed by considering various sets of data as distributions in their own right and simply outputting the frequency of different outcomes. This was useful when considering the data in its bulk. That is, the collection of all data regardless of the year published. The plots generated were more useful than simply gathering quantities related to the distributions as it gave a much greater appreciation for the shape of the distribution and if conclusions could be accurately drawn from it. Examples of histograms used on the project were those of the number of citations garnered by a paper and the career publication count of authors at WPI.

3 Data

With a solid plan to work towards the goals of the project in place it was time to implement it. Presented below is the information that was produced as a result of this methodology. Much of the data was time series data and it was found to be useful for later discussion to include markers indicating important dates in the history of WPI and the world. These markers were labelled by letters which translate to the following events:

a	Start of WWI
b	End of WWI
c	Start of WWII
d	End of WWII
e	Construction of Olin Hall

3.1 Departmental Makeup

The first piece of data that was studied with regards to the department itself was how the size of the department varied over time. A graph was created by counting the number of course catalogue entries in a given year (disregarding graduate students and Professors Emeritus) and plotting it against time. The resulting graph is displayed in Figure 1. As one would expect, the general trend is that the department has grown over the years. This is to be expected, because as WPI grew in size the physics department expanded with it. The one outlying period is a hump in the late 1960's and early 1970's. One interpretation of this could be that it was a result of the increased focus on fundamental science in the wake of the space race. When that interest shifted elsewhere after the fall of the Soviet Union the department contracted to its original growth trend.

The department has also undergone large changes in its overall makeup since its founding. Displayed in Figure 2 is a representation of the share of degrees held in the department. The set of qualifications that faculty members have held was divided up into categories representing Bachelor's level, master's level and doctoral level degrees. Then, for each year, the percentage of the department holding each level of degree was found and plotted against time. It was unexpected that up until a decade or two ago, a doctoral degree was not considered a requirement to teach and perform research at a university. In fact, there were no members holding a doctoral

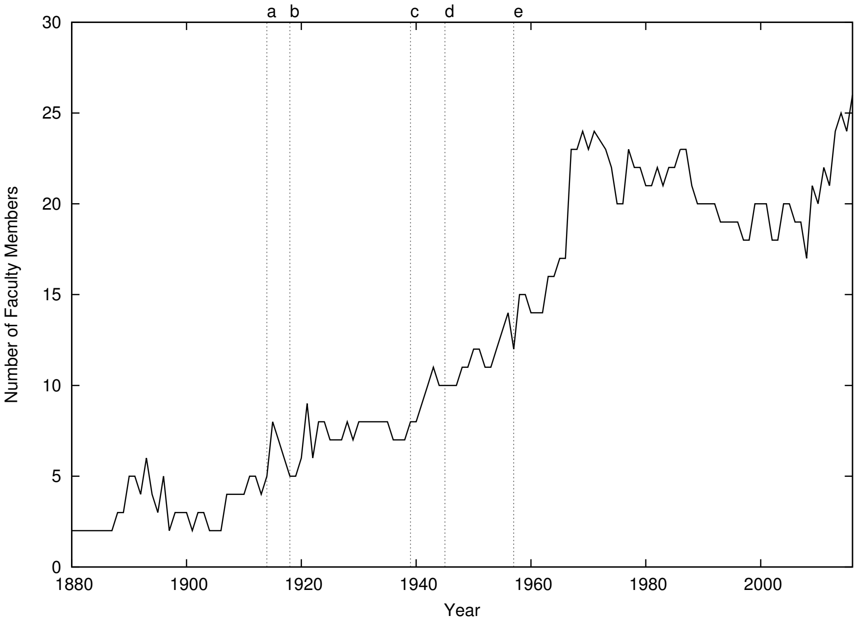


Figure 1: Annual Department Size

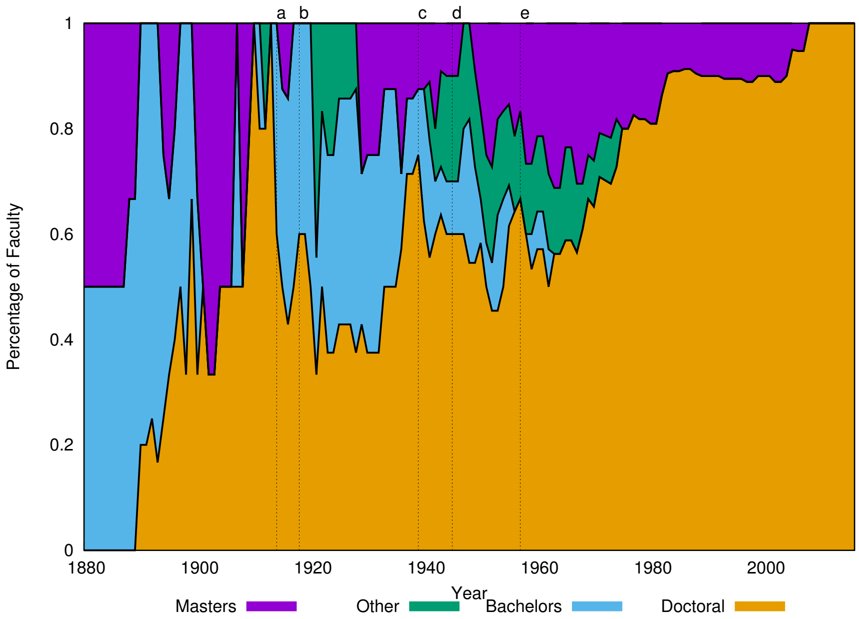


Figure 2: Annual Share of Degrees Held by WPI Faculty

degree for the first decade of the department.

Another change in the makeup of the department was the share of different positions that faculty held. Shown in Figure 3 is a plot of the relative frequency of different titles within the department. Just like in the plot of degree frequency the set of all titles ever held by WPI faculty was partitioned into three classes: Assistant Professor, Associate Professor, and Professor. All uncommon positions were categorized as "other." The relative frequency of each category was then tracked across the years by looking at entries in the course catalogues and plotted against time.

It was of interest to next study what the career of an average WPI professor is like. The first piece of information to be looked at was the tenure of faculty members at WPI. The set of unique members was retrieved from the database of course catalogue entries. Then the number of entries in the database that matched the name (and did not have the title "Professor Emeritus") was computed. A histogram of this data was created and is shown in

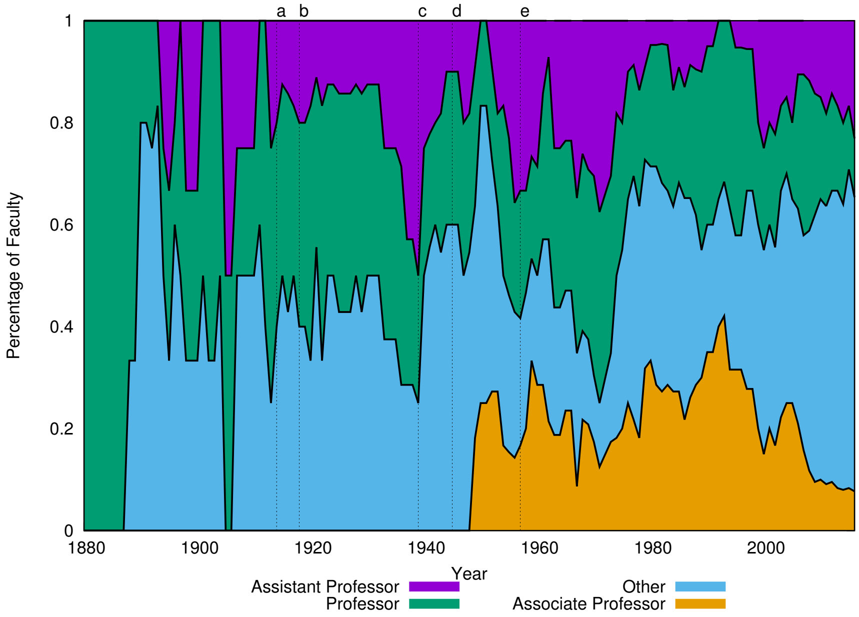


Figure 3: Annual Share of Titles Held by WPI Faculty

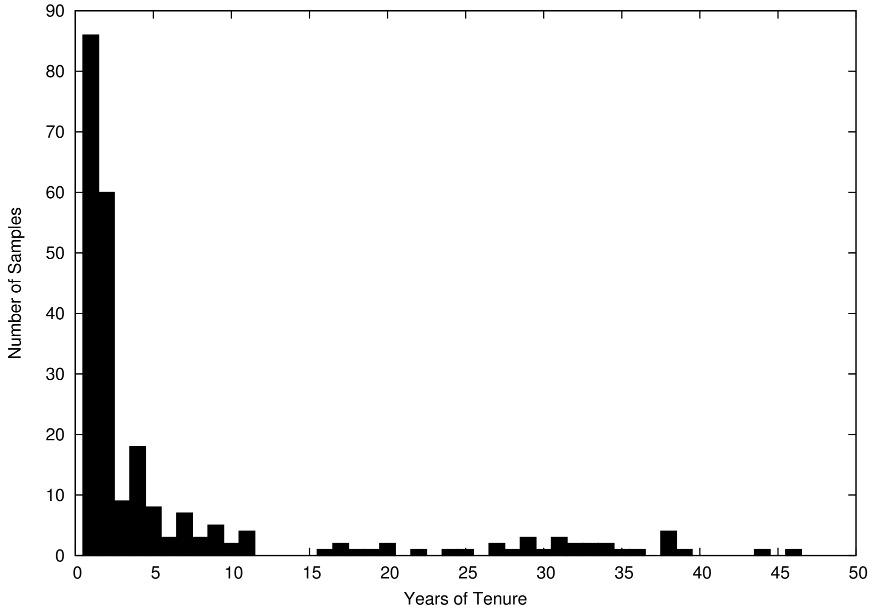


Figure 4: Relative Frequency of Career Lengths

Figure 4.

Another interesting facet of the careers of WPI faculty members was how often they were promoted or changed roles in the department. This was quantified by looking at the number of unique titles held by faculty members. For each unique name in the database, the number of unique titles in the database associated with the name was computed. This data was then turned into a histogram which is displayed in Figure 5.

Other information related to the careers of faculty members at WPI was tied to their publication tendencies. First was the number of publications made by a faculty member in his/her career at WPI. The list of unique names of faculty members was queried from the collection of course catalogue entries. Then, for each professor, a query was performed for publications containing that professor in the authors list. The name comparison method described in the Methodology section was used for author comparison. The number of results was collected and then displayed as a histogram which is

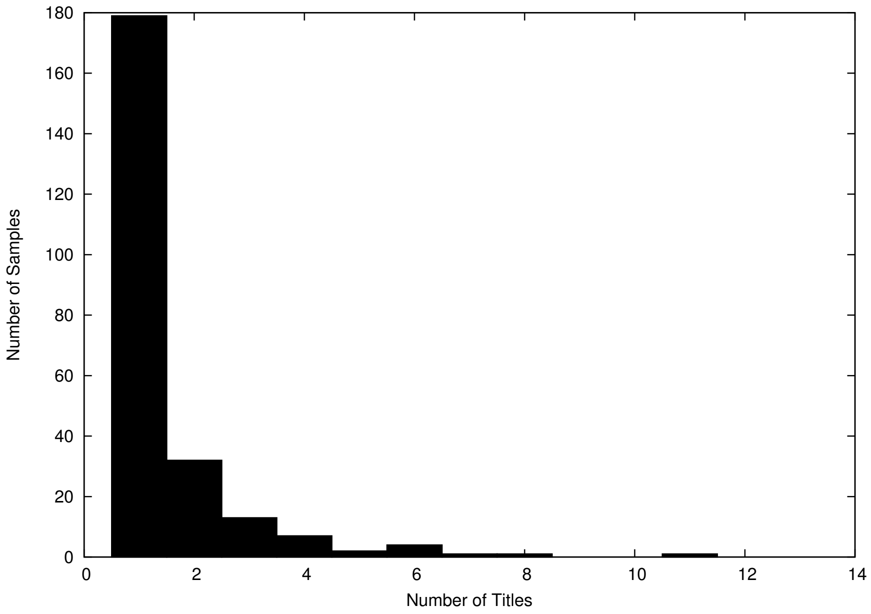


Figure 5: Relative Frequency of Number of Titles Held in Career

shown in Figure 6.

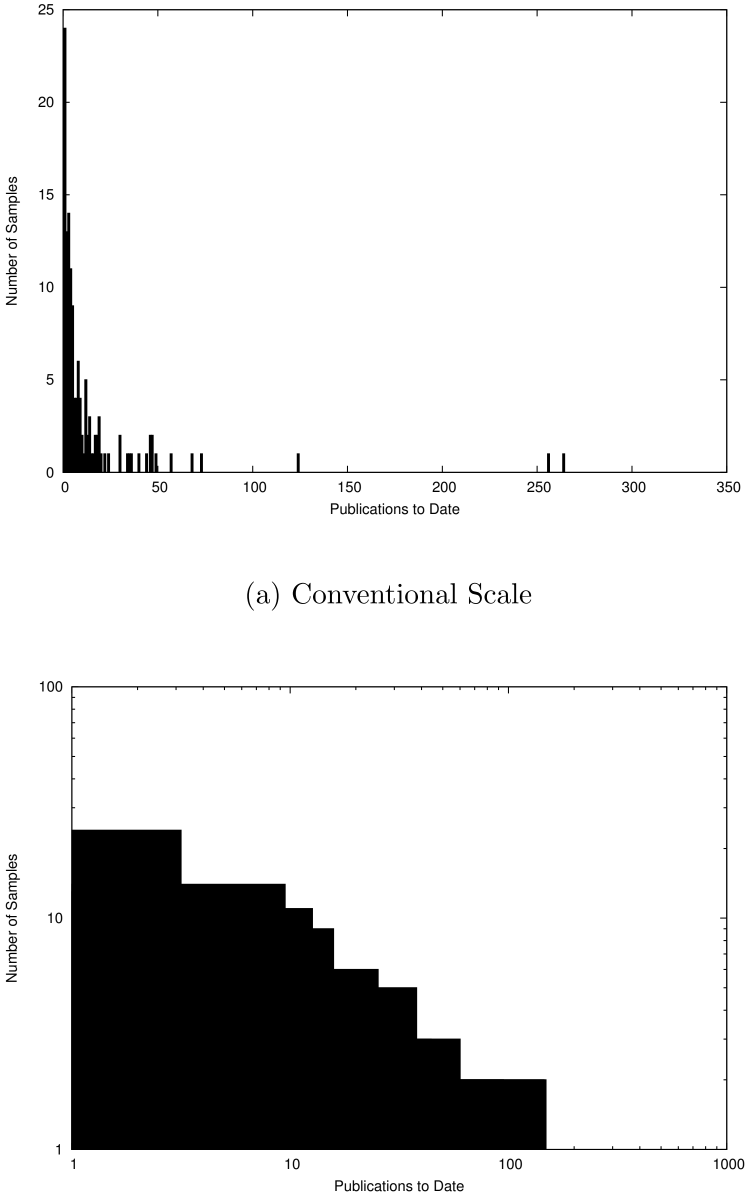
Also related to the publication tendencies of faculty members was where they published their research. Collected in the data obtained from the APS publication database was an indication of which journal the articles were published in. For each unique faculty member, the set of articles listing them as an author was queried from the publications dataset. This, again, used the name comparison method described in the methodology section. The number of unique journals articles were published in by that faculty member was then computed and plotted as a histogram which is displayed in Figure 7.

3.2 Academic Output

Next, information related to the publications made by WPI affiliated authors was studied for trends and patterns. The first piece of information to be looked at was the annual publication rate at WPI. That is, the number of publications made by WPI affiliated individuals on a yearly basis. The relevant information was pulled from the database and plotted to produce the graph shown in Figure 8. It can be seen that early in WPI's history, publications were fairly sparse until around 1960 when they began a steady increase which has not stopped yet.

Next, we looked at a measure of the influence of the papers published at WPI: the citation count. Each citation count was grouped into a distribution, the histogram of which is displayed in Figure 9. The power of a visual representation of data can be seen immediately in this example. The data neatly aligns itself to a Poisson-like distribution, giving us a clear intuition of how the citation counts are distributed in WPI's total publication history. This means that the majority of papers do not gather many citations, it is a very small amount that score above even 100 citations. As mentioned previously, the citation count reported by the APS is likely not accurate for most of the publications in this project's database. An accurate citation count requires the full digitization of records reaching back to the publication date of the resource in question..

The citation counts were also considered not in the bulk, but rather as time series data. This analysis was useful in understanding how the popularity of articles published at WPI has changed over the years. The average citation count for articles in a given year was plotted against time. The resulting plot is displayed in Figure 10a. A first impression upon looking at



(a) Conventional Scale

(b) Log-Log Scale

Figure 6: Relative Frequency of Career Publication Count

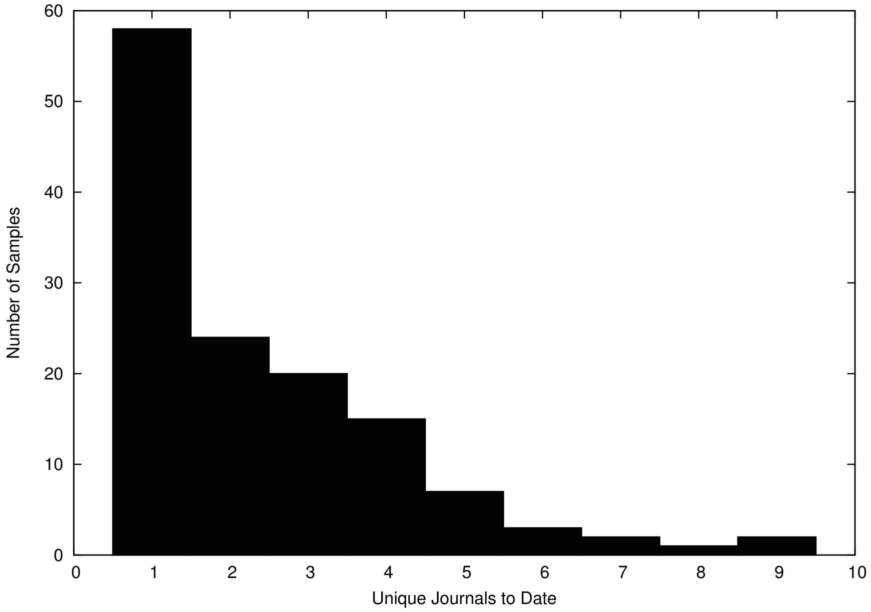


Figure 7: Relative Frequency of Number of Unique Journals in Career

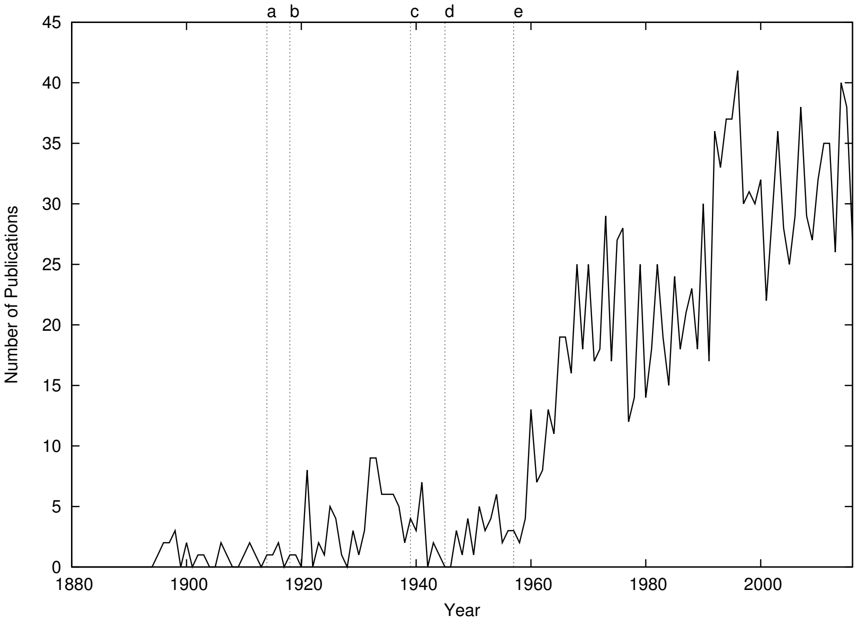
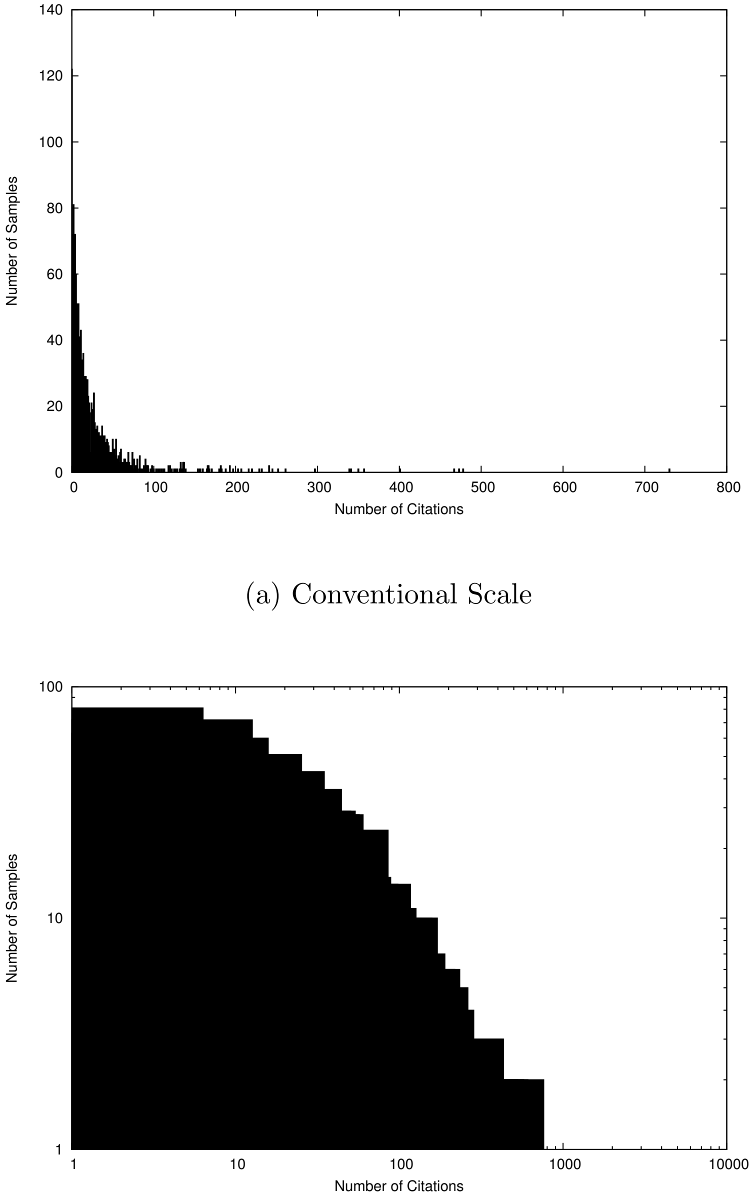


Figure 8: Annual Publication Count



(a) Conventional Scale

(b) Log-Log Scale

Figure 9: Relative Frequency of Citation Counts

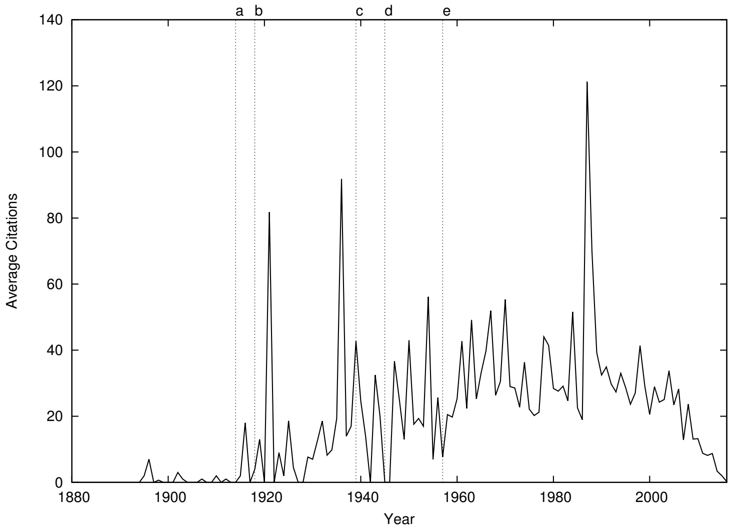
the graph is that it is fairly noisy. This is likely due to the relatively small sample size used in the study. There is however a general trend of increasing citation counts up to a point where it falls off. The increase in citations is likely due to the increase in the number of papers published in the field. The rapid fall of in average citations in recent years is likely not because of a deficit in the quality of papers, but rather that they have not had enough time to receive the same number of citations as their predecessors.

A better understanding of how citation counts were distributed among papers in a given year cohort was also sought. The standard deviation of the number of citations each paper from a given year received was calculated and plotted against time. The resulting plot is displayed in Figure 10b. Standard deviation is most easily understood as a measure of how spread out the samples in a distribution are. By this interpretation, the spread of the data follows the same trend as the average. This reinforces the original interpretation that recent publications have not had enough time to accrue similar numbers of citations as their predecessors.

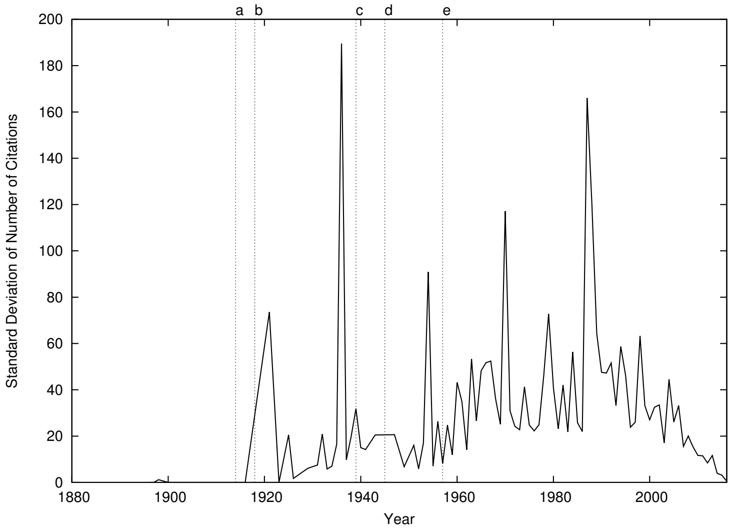
The next interesting type of data to be considered was the number of references made by authors at WPI. This is expected to be indicative of the field of physics in general and not of the academic culture at WPI. Reference counts were first aggregated in bulk and the relative frequency of each bin of samples was represented in a histogram. That histogram is displayed in Figure 11. As with the citation counts the samples of reference counts appear to form a Poisson-like distribution. This could be interpreted to say that there is some threshold of references a paper should have before it is considered rigorous. The total number of citations, however, is also limited by the scope of the paper which leads to the tail end of the distribution.

The number of references were then considered as time series data. Comparing the average number of references made in papers published in a given year against the year of publication produced the plot shown in Figure 12a. The trend shown in the plot is a general increase in the number of references made over time. This pattern has already been noted by other authors and is generally explained as resulting from the increase in the overall publications in the field of physics.[4] As more papers are available for authors to use, they will use more.

The standard deviation of the number of references made in publications written in a given year was also calculated. This value was plotted against the year in reference resulting in the plot shown in Figure 12b. The standard deviation showed a general increase over the years representing an increase



(a) Average



(b) Standard Deviation

Figure 10: Annual Moments of Citation Counts

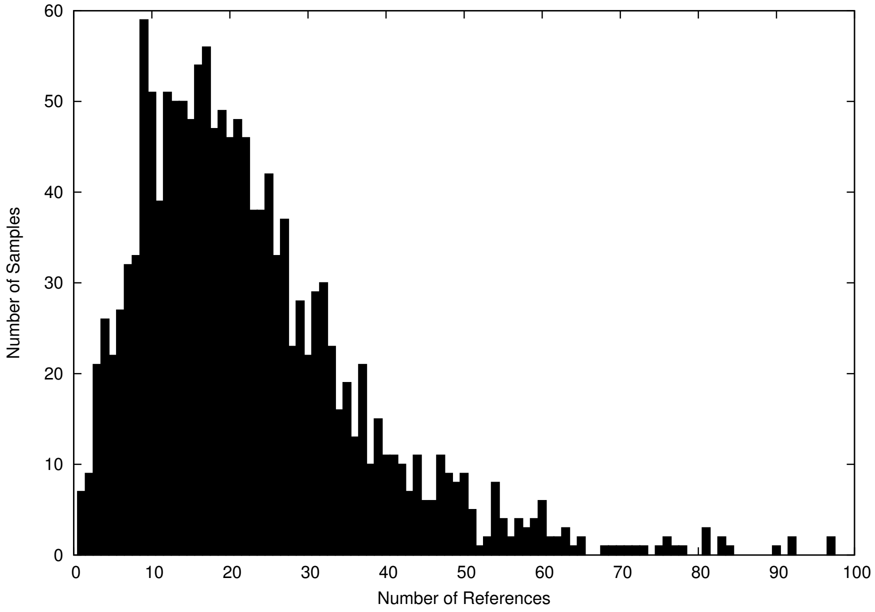
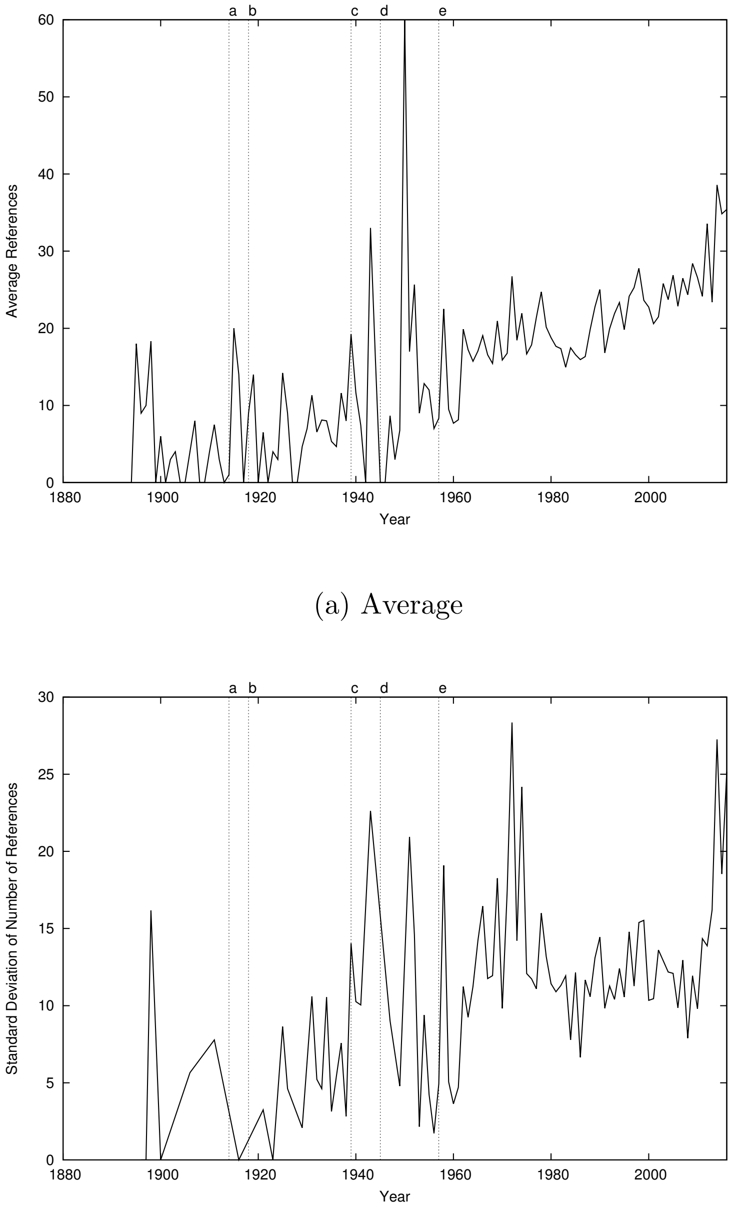


Figure 11: Relative Frequency of Reference Counts



(a) Average

(b) Standard Deviation

Figure 12: Annual Moments of Reference Counts

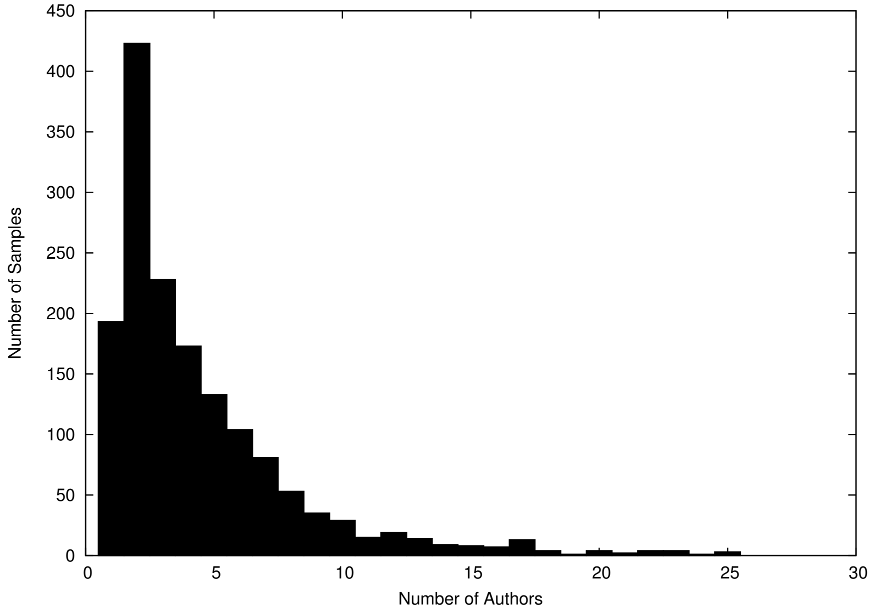


Figure 13: Relative Frequency of Author Counts

in the spread of samples in the dataset as time went on. One explanation for this increase is simply that the number of samples being studied increased over time. As more articles are published, a greater number of possible reference counts will be represented in the distribution yielding a larger standard deviation.

Another type of information that was considered during the analysis was the number of authors associated with a publication. This could be considered an indication of the amount of collaboration within and outside of the WPI physics department. As with other data, the number of authors associated with each paper was aggregated as a bulk distribution. This was then represented by a histogram which is shown in Figure 13. It is Poisson-like and suggests the interpretation that there is a normal number of authors when publishing a paper. There are however outliers in either direction represented.

The data was again split up into distributions by year and plotted as time series data. The result is shown in Figure 14a. Another well known trend in academic publication is the quickly increasing number of authors affiliated with single publications. The data in the APS database agreed with this reported trend as it showed a steady increase in the average number of authors associated with publications in a given year.

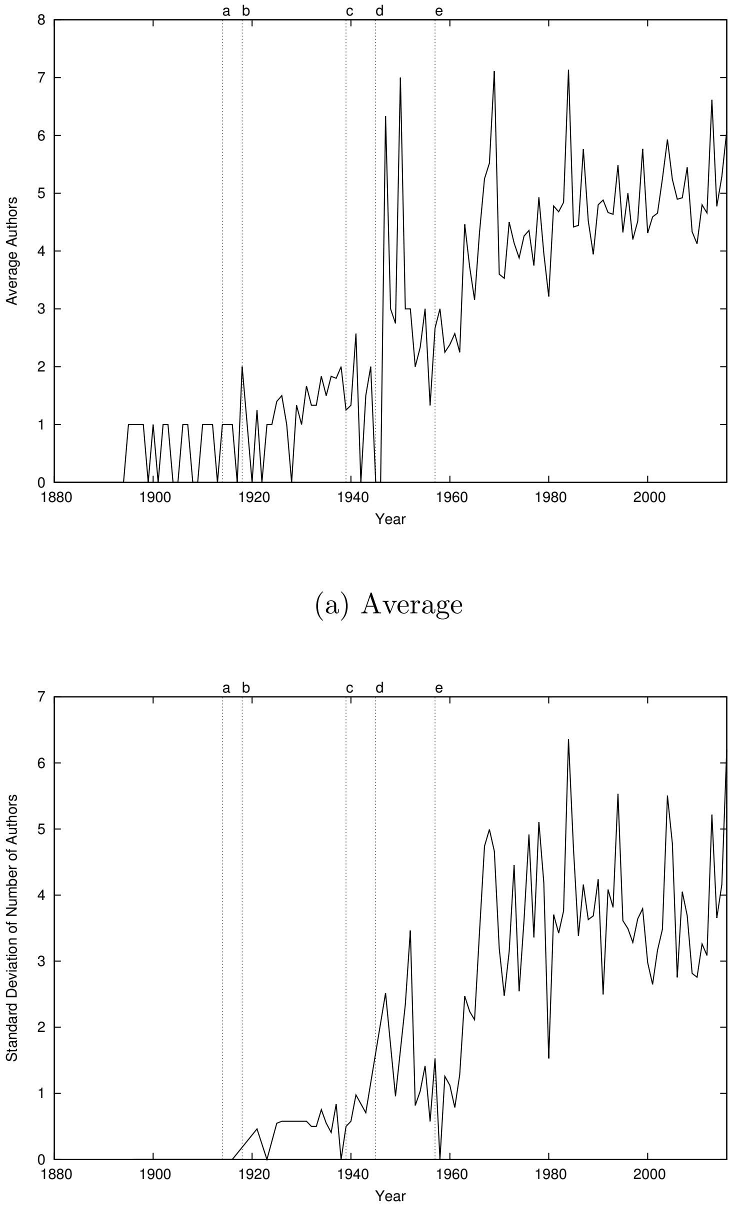
As with the previous distributions turned into time series data, the standard deviation was calculated as it changed over the years. This value was plotted against time and is represented in Figure 14b. Although the data is noisy, the standard deviation is again seen to increase steadily over time. This may be explained again as an increase in the number of samples as the years went along.

The next field to be looked at was the journal of publication. The set of publications made in each year was looked at and the percentage share of each journal was computed. This value was plotted against time on the percentage area chart shown in Figure 15. In the APS, journals are for the most part divided by the sub-field of physics that they focus on. By studying which journals WPI faculty published in, one can get an idea of what subfields were being studied at the university without the need for infeasibly complicated analysis.

The data shows a couple of different regions. Firstly the APS did not have more than one journal until the mid 20th century. At that point, the single journal was divided into five journals based upon the different subfields at the time. Once the journals split apart, however, one can see that the relative share of each journal stays approximately the same throughout time. This means that the broad research interests of the university haven't shifted fundamentally since at least the 1960s.

The final piece of information looked at was the fine grained date of publication of articles. Firstly, each article was given a number, from 0 to 365 representing the number of days into the year that the article was published. The average of this number was computed for the set of articles in each year and this value was plotted against time. The results are displayed in Figure 16. While it is noisy, the plot indicates that the average day of publication has not changed very much and is in the center of the year. This is what one would expect for a uniformly distributed set of samples.

A histogram was also produced to describe the relative frequency of publications made on the various days of the week. This is shown in Figure 17. The plot for the most part confirms the suspicion that the date of publication is uniformly distributed. There are, however, many days that appear



(a) Average

(b) Standard Deviation

Figure 14: Annual Moments of Author Counts

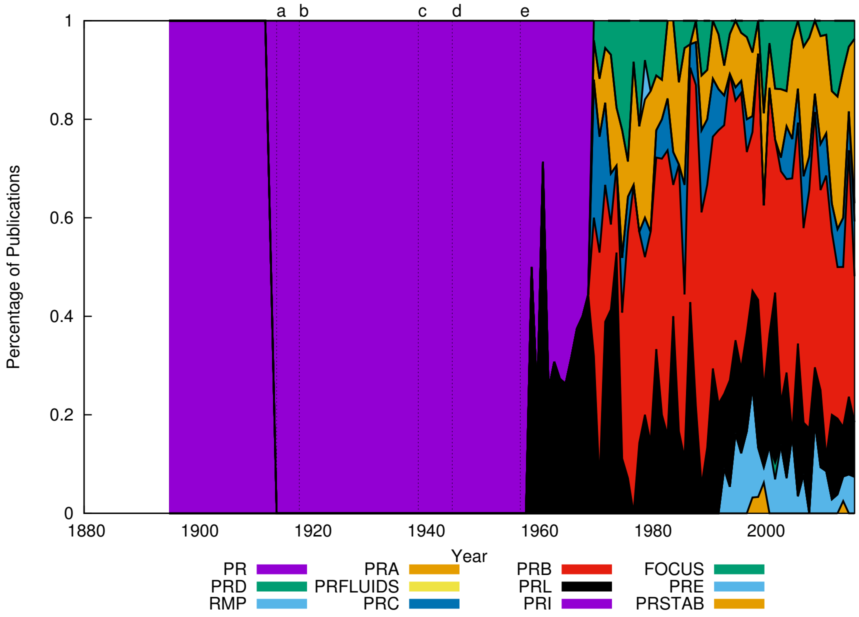


Figure 15: Annual Share of Publications by Journal

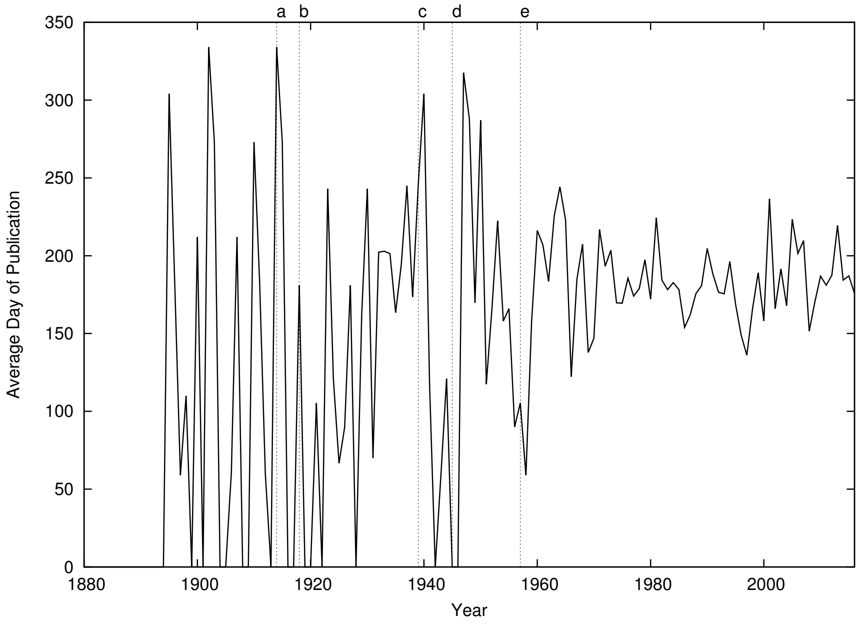


Figure 16: Annual Average Day of Publication

much more often in the dataset than others. Most of these correspond to the beginnings of months and weeks. It is likely that the APS, when not enough data is available, will return the article as having been published on the first of the month, which introduces this type of bias.

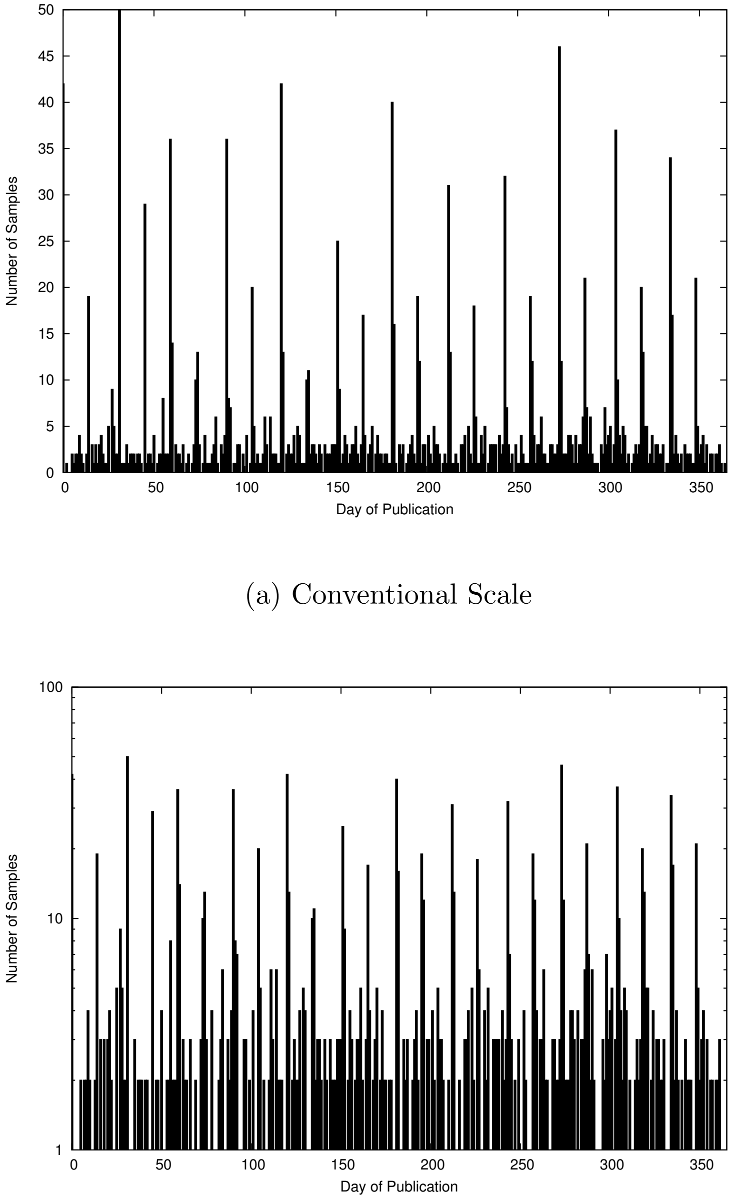
4 Discussion

The goal of this study, again, was to gain a better understanding of WPI's academic legacy. Now that we have determined the patterns and trends in the data collected during the first stage of the project it is important to provide realistic interpretations of these trends. These interpretations will help give a clue as to why the patterns exist and if they can be used to help achieve our goal. The below featured discussion is a survey of the most significant relations between data that were seen while working on the project.

4.1 Important Trends

One of the most surprising discoveries found during the project was that it was only very recently that a doctoral degree (or equivalent) was required to work in an academic department. The data in Figure 2 shows that the department began with an even split of individuals with bachelors and masters degrees. Up until 1970 or so, the degrees didn't seem to show any strong trends. The share of individuals with doctoral degrees did increase at times, but it also dropped by nearly the same amount at other times. After 1970, however there is a strong increase in the number of people holding a doctoral degree.

The American Institute of Physics (AIP) is an organization focused on improving the study of physics in the United States. They fortunately have a repository of data related to national level changes in field of physics. In their publication "Trends in Physics PhDs"[5] they show data which suggests a national increase in the number of people seeking and earning doctoral degrees beginning in 1950. According to the graph shown in Figure 1 the department size stayed approximately the same in the years 1950-1960. What likely caused the increase in the share of individuals in the physics department holding doctoral degrees wasn't pressure from WPI, but rather a larger change in the field. The reason it was delayed from the initial onset of PhDs nationally was because WPI wasn't hiring many faculty members



(a) Conventional Scale

(b) Semi-Log Scale

Figure 17: Relative Frequency of Day of Publication

until almost a decade later.

Another important trend was the increase in the average of citation count of papers published at WPI over the years. Figure 10a shows a slow rise in the average citation count beginning in around 1930. This is followed by a sharp decline in the average citation count starting in 2000. A decline is fairly easy to explain as being caused by the fact that the publications have not had enough to to accrue citations. It takes a while for a publication to gain recognition and for other researchers to make use of it. The increase in citation counts could have a number of contributing factors.

A report by Redner[6] provides a history of global citation trends specific to the journal "Physical Review." The study was focused on a 110 year period of publications and the factors affecting their citation counts. Page 3 shows that the total citation count of all publications in Physical Review has increased exponentially over time. This provides support for the trend being global and not associated with the WPI physics department in particular. A large increase in the number of individuals studying physics as well as the increased availability of articles through improvements in technology are typically cited reasons for this rise in citation counts.

Another interesting point made by Redner is that it is possible for an article to be published on a subject that becomes popular many years later. This appears as a paper that has a low citation count for a long time until a particular year when it rises quickly. Five popular papers showing this feature are shown in Figure 8 of Redner. This is an additional, although likely less significant, reasons for the decrease in average citation counts beginning in 2000. That is, publications may have been produced on subjects that are not yet popular.

4.2 The Typical WPI Faculty Member

An interesting way to look at the data is to ask the question "what is the typical faculty member at WPI like?" Questions like this help ground the data in reality. By providing a comparison to real life experiences one can better place the data against their own experiences and help determine if it agrees with them or not. Putting data in these terms also helps to reveal the hidden assumptions people make about the physics department that may not be true. These will both benefit the goals of this project.

Firstly, the average tenure of a WPI physics faculty member is 6.6942 years. Looking at the histogram in Figure 4 reveals that there are really two groups of faculty members. Long term faculty who have an average career length of closer to 30 years and short term faculty who only stay around for a couple of years. The average faculty member will be in one of these two groups.

The plot in Figure 5 shows the relative frequency of faculty who have held differing numbers of titles in their career at WPI. This is an indication of the culture of the business side of WPI and how frequently promotions occur. Computing the average of the distribution gives a value of 1.5416 titles implying that most faculty members don't change their title here. This was surprising, but makes sense when thought about. One typically thinks of the department in terms of its tenure track faculty who will typically move between a couple of different titles on their course to a full professorship. However, there are also many associate, assistant and adjunct professors who join the department only to leave for other career opportunities in a few years.

Also important is the number of publications a faculty member makes in their career. The relative frequency of career publication count is shown in Figure 6 and is Poisson-like. The average of the distribution may be computed to be 15.3485 publications. While this number seems a little low, this likely caused by the fact that there are many faculty members in the WPI physics department who do not produce publications in their time here. These are likely the adjunct or associate professors who's primary responsibility is teaching.

One may also try to judge the impact of the average faculty member by looking at the career citation counts for faculty members at WPI. As this section focuses on the subset of faculty who produce research, an average was taken over those who have publications. This average was found to be 456 citations in a career. One source gives the average citation rate of papers in physics as 15 [4] implying that the papers coming from our average faculty are of higher popularity than those of his peers in the field.

Similar to citations are the number of references made by the physics faculty at WPI. Again, an average was taken over the subset of faculty members that produce publications. The value of this average was 324 references per faculty career. This is approximately the value of the average publications in a WPI physics faculty career times the average references in a paper from WPI. This indicates that there is no complex grouping of samples in with regards to the number of references made.

4.3 Reactions to Historical Events

Placing changes related to the WPI physics department in the context of historical events is another important way to look at the data in this project. This additional view has the potential to lead to useful relations and interpretations that would be otherwise missed. It may also provide useful examples of the WPI physics department reacting to historical events that manifest themselves again in the future. These may be used as case studies for future researchers to take lessons from. Finally, the additional context will help provide a better understanding of WPI's involvement in physics, therefore furthering the overall goals of the project.

The two periods of history with the most significant effect on WPI would have likely been World War 1 and World War 2. Most academic pursuits were put on hold and the university was mobilized for the production of military goods due to its machine shops. One would expect that this slowdown in research would manifest itself by a drop in the publication rate during those years. Referencing Figure 8 reveals that there is a drop during World War 2.[1] The publication rate during World War 1 does not seem to change significantly, however there are so few samples in that time period that it is unlikely to be significant.

While digitizing information from the course catalogues it was noticed that two atomic physicists at WPI, Karl Wilhelm Meissner and Robert Thompson Young, Jr, had leaves of absence that overlapped with World War 2. During that period of history, the Manhattan Project was working towards the development of an atomic bomb by the allied forces. Many physicists from around the United States were recruited for the effort and it seemed plausible that they could be involved in the project. The Lab Historian at Los Alamos National Labs was contacted and unfortunately could find no records of the individuals working at the labs during the Manhattan Project. He did however indicate that the Los Alamos staff only made up 1%-2% of the total workforce of the Manhattan Project and that they may have been involved in other capacities.

5 Concluding Remarks

5.1 WPI's Academic Legacy

A general appreciation for the contribution WPI has made to research in physics over the years may be gained simply from the data gathered during this project. However, out of the bulk of the publications occurring here, two happen to stand out in particular. These are papers by Albert W. Hull and Richard A. Beth. Hull provided the theoretical background to describe the motion of electrons in a specially configured system of charged plates in a magnetic field. [7] The importance of the work was that it provided a method of measuring the electron's charge to mass ratio e/m and provided the most accurate value for that fundamental constant at the time.

Beth's work was similar to Hull's in that it involved the experimental determination of fundamental values. [8] This time, the photon's angular momentum was measured to a high degree of precision by studying the torque caused by light striking a fine quartz fiber. The paper ended up being the single most cited work in the entire database with 478 unique citations at the time of this writing.

6 Future Work

Although our goals were accomplished there were many areas that could not be fully explored due to a lack of time or resources or both. A number of future projects could be performed to build off of the work that has been laid out in this paper. Presented below is a discussion of the significant areas where future researchers may want to pick up. This section has been divided into two major areas of focus. One dealing with topics that have to do with the data sources used in this paper and another having to do with potential changes to the methodology used in this paper.

6.1 Further Data Sources to Explore

The relatively small number of sources of information is one area that could be improved greatly by future researchers. By limiting the study to WPI's course catalogues and to the APS database, not all publications made by authors affiliated with WPI could be captured. This is a problem as it limits the reliability of any findings this study has made. It would be a useful project to attempt to form an exhaustive collection of information on all of the articles ever written by faculty associated with WPI. One would then perform a similar analysis to the one in this study and see if the results are reproduced with the expanded dataset.

Work could also be performed on increasing the scope of information collected on faculty members at WPI. The institution, as a business, needs to maintain records of some type on the people that they employ. If these have survived over the years, then they could provide a potentially fruitful source of information for investigating the department. As an example, it was mentioned that there were two faculty members that could have been involved with the Manhattan Project. A personnel file on them may lend insight into what they were doing on their leaves of absence and help clear up this mystery.

6.2 Improvements to Methodology

While the methodology used in this project was expanded as far as it could in the time allocated, there were limits to what could be done. Certain types of data could not be reasonably gathered by the project. Immediately coming to mind are the street addresses of faculty members at WPI and full text copies of publications in the APS database. The first could be copied with enough effort or with the development of an automated system. The second, however would require permission from the APS and may not provide significantly more data than what is already available.

One interesting line of research that could not be explored in the time available for the project would be looking for correlations between different attributes in the collected data. For instance producing scatter plots of career citation counts for faculty members against attributes such as tenure, career publication count, or career reference count. One could then look for trends indicating a potential correlation between two attributes. This method of investigation has the potential to yield valuable insight into who becomes a successful researcher in the WPI and who does not. These results could then be interpreted into lessons for future researchers to learn from.

Module Name	Version
MySQLdb	1.2.5
mechanize	0.2.5
json	2.5.1
socks	1.5.7
numpy	1.10.4
statistics	1.0.3.5
nameparser	0.5.1

Table 1: Python Module Versions

Appendices

A Software

The software produced during this project may be accessed as supplemental material to this paper. All software was developed for python version 2.7 interpreter and interfaces with MariaDB version 5.6. The versions of python modules used in the project are listed in Table 1.

B Datasets

The datasets collected in this project may be accessed as supplemental material to this paper. They are formatted as the output of the application mysqldump. Mysqldump produces a SQL formatted dump of the entire database in a plain text file which may be easily imported into mysql or, with minor modification to the output file, any database management system that supports the SQL standard. The database includes the tables:

aps_article_metadata Includes all single valued parameters of publications collected from the APS database.

aps_author_association Contains pairs of authors and IDs to the articles they are associated with. By querying all authors associated with one paper from this table, the author list may be retrieved.

aps_query_history A list of the queries performed on the APS database and if the query is considered valid or not.

aps_references Similar to the table of authors, includes pairs of references and articles. The reference list of an article may be retrieved by querying the references associated with it from this table.

course_catalog_faculty_entries The raw entries collected from WPI course catalogues.

References

[1] M. M. Tymeson. *Two Towers*. Barre Publishers, Barre, Massachusetts, 1 edition, 1965.

[2] Herbert Foster Taylor. *Seventy Years of the Worcester Polytechnic Institute*. Worcester Polytechnic Institute, 1937.

[3] J. Agar. *Science in the Twentieth Century and Beyond*. Polity Press, Cambridge, 1 edition, 2012.

[4] S. Redner. How popular is your paper? an empirical study of the citation distribution. *The european physical journal b*, 4:131–134, 1998.

[5] P. J. Mulvey and S. Nicholson. Trends in physics phds. *Reports on enrollment and degrees*, 2014.

[6] S. Redner. Citation statistics from more than a century of physical review. *Arxiv*, 2004.

[7] A. W. Hull. The effect of a uniform magnetic field on the motion of electrons between coaxial cylinders. *Physical Review*, 18(31), 1921.

[8] R. Beth. Mechanical detection and measurement of the angular momentum of light. *Physical Review*, 50(115), 1936.