# ProQuest Dissertation Analysis

Kishor Patel[1], Sergio Govoni[1], Ashwini Athavale[1], Robert P. Light[2], Katy Börner[2]

[1]Kishor.Patel@proquest.com, Sergio.Govoni@proquest.com, Ashwini.Athavale@proquest.com
ProQuest LLC, 7500 Old Georgetown Road, Suite 1400
Bethesda, MD 20814, USA

[2]katy@indiana.edu, lightr@indiana.edu
CNS, SOIC, Indiana University, 1320 E. Tenth Street
Bloomington, IN 47405, USA

## Introduction

Productivity measurement has become a major issue for university leaders. Federal and state governments support teaching and research with significant investments. When university leaders are seeking new funding, it is not uncommon that they need to justify their request with productivity measurement metrics and equally important research output consumption metrics. However, it is often very difficult for university leaders to generate these metrics as they lack access to relevant data and tools to analyse and visualize large amounts of data.

Equally important for university leadership is strategic foresight, i.e., knowledge on emerging new research areas or strategic changes in the portfolios of peer institutions (Lebo, 2011). While some of this knowledge can be gained via social and professional networking, a global, worldwide and up-to-date perspective benefits from the systematic analysis of relevant datasets. Ultimately, university leaders need means to continually monitor and constantly improve their graduate programs and topical expertise profiles in response to changes in the scientific landscape.

Interested to address the diverse needs of university leaders, ProQuest analysed its ProQuest Dissertation & Theses Global (PQDT Global) database, an extensive and trusted collection of 3.8 million graduate study dissertations with 1.7 million full text records and editorially assigned metadata created by subject area experts. The database offers comprehensive North American and significant international coverage. Worldwide access to the database is logged at the dissertation level by ProQuest. Usage data mining is important for understanding user behaviour (Srivastava, Cooley, Deshpande, Tan, 2000). The ProQuest Dissertations Dashboard released in 2014 provides easy access to dissertations, metadata, and usage data. It is available for free to leaders of any university that shares dissertation data with ProQuest.

The remainder of the paper is organized as follows. The next section discusses data acquisition and preparation. Subsequently, five different studies are detailed that answer questions of particular interest to university decision makers. The final section discusses proposed and planned improvements of ProQuest dissertation data in support of future analyses and visualizations.

## Data Acquisition and Preparation

The PQDT Global database is an official offsite repository for dissertations and theses acknowledged by The Library of Congress. As of December 2014, the database comprised metadata records for 3.8 million documents with 1.7 million in full-text PDF format . A two-level subject category classification with 11 primary and 411 secondary classes is used to organize dissertations topically. Usage data, i.e., information on which dissertation was downloaded from where how often, is available from 2012 onward.

In this joint project, a large subset of PQDT Global data was provided to Indiana University in the form of 2,946 zipped XML files. The files were parsed via a Python script into a SQL database powered by PostgreSQL 8.4. The final dataset comprises 2,894,414 distinct doctoral dissertations and master's theses records published from 1637 to 2013 with the majority of records published between 1960 and 2012. Nine of the top ten schools are based in the United States, as are nearly 80% of all records. Most other records are from North America and Europe, while data from other regions are notably absent. Advisor data exists for nearly half of all the records making it possible to study scholarly genealogies (Sugimoto, 2012; Ni & Sugimoto, 2012)

## ProQuest Data Analysis and Visualization

Analyses were conducted and results visualized to answer questions that seemed of particular interest to university leaders and those seeking to assess the performance of a school as a whole. Exemplarily, the results of five studies are presented here.

***Study 1: How much attention are my school's dissertations getting?***

A school's ability to generate interest in their students' dissertations may not only reflect the reputation of the school, but have long-term effects on those students' marketability and also in attracting future generations of students to join the school.

Usage data for dissertations for a specific set of (peer) institutions and a selected subject area can be plotted and compared to answer this question. Exemplarily, Figure 1 plots the production and access data for computer science dissertations for a selected institution given in red and labelled 'Subject University' and two groups of peer institutions rendered in green and blue. Other institutions that have published computer science dissertations are given in grey. The three institutions in the top-right corner of the plot—publishing many theses that attract many views—include both well-regarded private research institutions as well as for-profit colleges with practically open admissions. This implies that while thesis production and usage are important, they should not be used as a sole indicator for the quality of a program.

***Study 2: Are my degree programs' reputations growing stronger or fading over time?***

As download activity data is available from 2012 onward, trends over time can be studied for specific sets of institutions and subject areas. For example, Figure 2 shows access counts for sociology dissertations for a 'Subject University' and seven peer institutions labelled A to G that act as a pre-defined control group. As the number of accesses depends strongly on the number of dissertations per institution (and ultimately also the number of PhD students and faculty members), normalization becomes critical. In order to normalize for the size of the sociology school at that institution, the number of sociology dissertations published at each institution since 2007

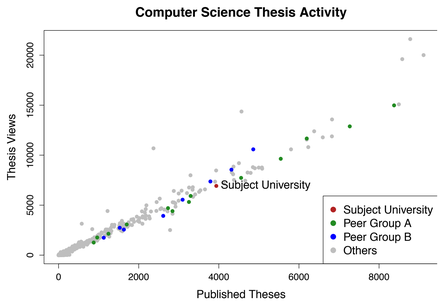


Figure 1. Comparing Subject-Area Specific Thesis Access Activity with Peer Groups

was computed. Figure 3 shows the same access data normalized by this number. While in Figure 2, the thick line for Peer B is steady, with an expected depression through the summer months and in December, it seems to be consistently below the group average. When adjusted to account for the number of dissertations published since 2007 (arbitrarily chosen), however, a major change can be observed (see Figure 3). Basically, while Peer B publishes relatively few dissertations in the realm of sociology, those that they do publish see very heavy use. Rather than a perennial underperformer, this paints the picture of a small, but very well-regarded sociology department.

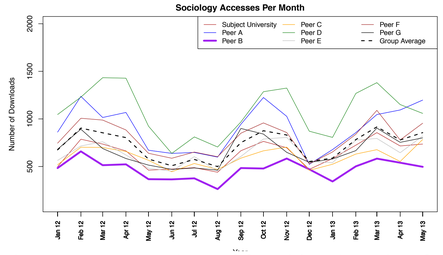


Figure 2: Raw Access over Time Graph for Sociology Dissertations of a Peer Group

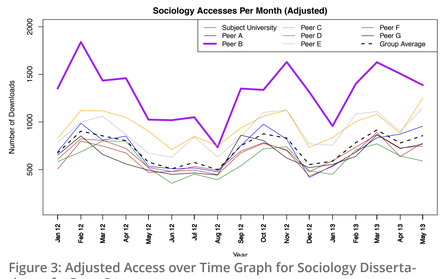


Figure 3: Adjusted Access over Time Graph for Sociology Dissertations of a Peer Group

***Study 3: How can I quickly compare the number of dissertations and associated download activity for a large number of universities?***

Given all dissertations or dissertations in a certain subject area, university leaders might like to understand the "market share" of an institution within a comparison or peer group. Here, a treemap visualization (Shneiderman, 1992) can be used to plot the number of dissertations and access counts providing a global overview of activity for a certain time span. A treemap is a space-filling data visualization (see Figure 4). Given an area, a recursive subdivision is used to lay out a tree structure (e.g., a dataset that consists of two peer groups labelled A and B, each with a set of different peer institutions) without producing holes or overlaps. Area sizes may correspond to the attributes of the subtrees they represent and they may be labelled and color-coded.

In Figure 4, two peer groups of institutions are compared. Each institution is represented by a rectangle. Each rectangle is sized based on the total corpus of computer science dissertations available in the ProQuest dataset for that institution. Colours tell how frequently the average dissertation at that institution is accessed in comparison to the group average. Computer science dissertations written at Universities L, O, and R are accessed more frequently than the group average, while those published at Universities G or P are accessed less frequently.

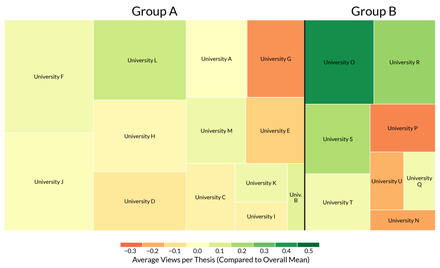


Figure 4: Treemap Comparing Thesis Production and Usage in Computer Science Within Two Peer Groups

Treemap visualizations can also be used to provide an aggregate view of a university's doctoral dissertation publication activity by primary and secondary subject categories for a specified time period, see Figure 5. Rendering the same visualization for peer institutions supports the comparison of subject area strengths at a more detailed level.

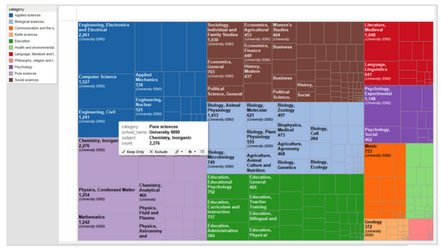

Figure 5: Treemap Showing Dissertations by Subject Area for a Particular University

***Study 4: How is dissertation information flowing in and out of my university?***

Universities are both producers and consumers of information (Mazloumian et al., 2013). Administrators are interested to understand which dissertations from which universities are used at their own institution but they also want to know who is accessing their own institution's dissertations. Plus, they might like to compare this in-flow and out-flow of information with the flows calculated for other universities.

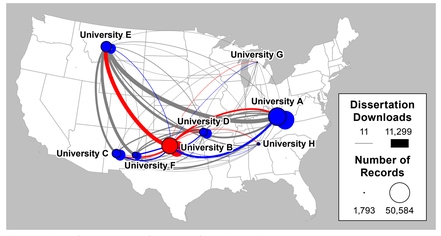


Figure 6. Information Flows within Peer Group

The example in Figure 6 looks at information flow between a group of peer schools. One institution, labelled University B, is highlighted. Red edges depict information flowing out of that institution, while blue flows show information flowing into that institution. The thicker the line, the greater is the number of dissertations. (Information always flows clockwise on the curved lines). At a glance, it is clear that while University B frequently consumes its own dissertations, it is not using other institutions' dissertations as often. On the other hand, University E is a heavy consumer of dissertations from other members of the peer group. A detailed listing of in, out and self-usage flows can be tabulated, (see Table 1).

Table 1: Total Information Flow for Schools in a Peer Group

| Institution | Outflow | Inflow | Self-Usage |
|---|---|---|---|
| University A | 14,083 | 8,182 | 11,299 |
| University B | 12,293 | 8,073 | 8,680 |
| University C | 8,082 | 4,489 | 8,058 |
| University D | 8,385 | 8,087 | 6,991 |
| University E | 2,813 | 22,363 | 5,207 |
| University F | 10,062 | 3,992 | 5,797 |
| University G | 440 | 1,234 | 292 |
| University H | 2,521 | 2,259 | 732 |

***Study 5: How is science growing and changing?***

The scientific landscape is evolving continuously. Similarly, the expertise profile and the reputation of universities is changing over time as faculty is hired or retires, funding starts or ends, and new priorities are established. Figure 7 shows the top ten secondary subject areas in the ProQuest database in terms of dissertation production from 1960 to 2013. Horizontal lines mark the maximum number of dissertations published in each subject area.

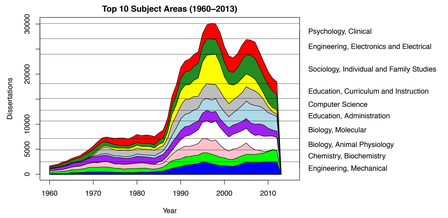


Figure 7: Top-Ten Secondary Subject Areas for Thesis Production (1960-2013)

Computer science (in light blue) emerges starting in the 1960s, while molecular biology (in white) first shows in the 1980s. Meanwhile, the boom in sociology (in yellow) enjoyed in the 1990s has faded and animal physiology (in pink) has suddenly vanished with the coming of the new decade. The latter may actually be a change in terminology or classification, rather than a shift in scientific focus. By drawing this graph as a whole and comparing it to the output of an individual university, a leader can understand where the institution is as part of the scientific community as well as see the effect of long-term policies such as major funding initiatives in a field or the establishment of strategic priorities.

## Discussion

This final section discusses proposed and planned improvements of the ProQuest dissertation data in support of future analyses and visualizations.

***Identifier alignment:*** As with other data sets, PQDT Global presents identifier alignment challenges. Problems arise in two ways: The first is the lack of synchronization between the dissertation data repository and the usage data repository. Each has its own independent set of IDs for the data. Providing the same unique IDs for both datasets will reduce errors.

***Unique institution names:*** The second issue is that of identifying unique universities, both as sources and consumers of data. Regional campuses and specific schools are often assigned their own school codes as producers or accounts as consumers. While there may be business requirements for maintaining multiple classifications, they can make determining the full output or ingest of a university difficult. As the two identification systems are completely separate, it is difficult to link output and ingest. ProQuest is working on a data normalization that addresses this problem.

In this project, the research team manually harmonized identifiers for a small group of universities. To make full use of this data on a global basis, a thorough effort will be needed to determine what constitutes a singular entity and to assign all accounts and school codes to the appropriate entity. This will certainly be a complex task, as there are regional campuses to account for and non-university entities such as corporations that consume material. If it is not handled proactively, then it must be addressed on a case-by-case basis, which is often more time-consuming and can lead to inconsistent implementation methodology.

Note that determining the appropriate level of aggregation is non-trivial. For example, it might make sense to collapse Indiana University (IU) with ZIP code 47401 and IU with zip code 47405 as both are in Bloomington, IN. However, collapsing all eight IU campuses across the state of Indiana might also be valid in some circumstances. Obviously, the higher the aggregation, the more likely it will be that IU makes it into top-n lists. A compromise between maintaining geographic identity and acknowledging the work of an entire university system is needed.

***Subject Areas:*** Of the 11 primary category and the 411 secondary subject areas, the top 20, shown in Table 2, account for 27.5% of all subject area assignments.

***Normalization:*** During the process of completing this project, there were a great many questions about normalization of data. This is not unexpected, as this is an active area of study in library and information science. However, as illustrated by Study 2, normalization is critical to understanding what the data is truly saying. By and large, normalization methods must be matched to the individual question being addressed.

Normalization for universities can take place in a number of ways. One is the creation of a peer group. This pre-defined group should be nominally similar in most attributes, allowing for those notable differences in characteristics to show without being lost in the noise of trying to compare to much larger and smaller institutions. Comparison to the entire university landscape grows more difficult, as it is challenging to draw a meaningful comparison between a major state university and a small hometown college. In this case, normalization factors can be applied. Some of the values that can be normalized against are faculty size (as a whole or within a target department), enrolment (again, within the university or the department), endowment, or number of theses generated over a number of years. This does raise a number of issues, though. One is the availability of such data. While some of this information is publically available, at least at the university level, gathering and cleaning it is a non-trivial task. The other is that not everything scales with size, and factors that do may not scale in a linear fashion. Having twice as many faculty or endowment dollars in the Computer Science Department as in the Chemistry Department does not automatically mean that twice as many computer science dissertations will be generated. Normalization is a means to help make things that would otherwise be hard to compare reasonably similar, but it is impossible to completely account for the differences between the large and the small with any simple metric.

Table 2: Top 20 Subject Areas by Number of Times Assigned (All Time)

| Rk | Class Code | Subject Description | Theses |
|---|---|---|---|
| 1 | 622 | Psychology, Clinical | 97279 |
| 2 | 544 | Engineering, Electronics and Electrical | 95385 |
| 3 | 628 | Sociology, Individual and Family Studies | 88809 |
| 4 | 727 | Education, Curriculum and Instruction | 81946 |
| 5 | 514 | Education, Administration | 77362 |
| 6 | 984 | Computer Science | 75292 |
| 7 | 433 | Biology, Animal Physiology | 70964 |
| 8 | 307 | Biology, Molecular | 70823 |
| 9 | 487 | Chemistry, Biochemistry | 67888 |
| 10 | 548 | Engineering, Mechanical | 62280 |
| 11 | 405 | Mathematics | 58626 |
| 12 | 631 | Sociology, Ethnic and Racial Studies | 53223 |
| 13 | 623 | Psychology, Experimental | 52058 |
| 14 | 525 | Education, Educational Psychology | 51764 |
| 15 | 515 | Education, General | 51571 |
| 16 | 490 | Chemistry, Organic | 47853 |
| 17 | 451 | Psychology, Social | 47847 |
| 18 | 453 | Women's Studies | 46062 |
| 19 | 615 | Political Science, General | 44972 |
| 20 | 410 | Biology, Microbiology | 44752 |

***Unique author names:*** A more advanced, but more challenging goal is author name disambiguation. While dissertations are by nature single author, many records include information on advisors and committee members. It might make a compelling case to be able to show the ongoing work on a dissertation writer in the 1970s as an advisor to students in the decades since. While global identifiers such as ORCID and efforts such as ScienceCV (http://www.ncbi.nlm.nih.gov/sciencv) are still in the process of gaining acceptance, they may prove helpful in creating these linkages.

***Future directions:*** Currently, ProQuest dissertation data is not linked to publication, funding or other data. However, there is much interest in being able to study career trajectories in a more comprehensive manner (Ni & Sugimoto, 2012; Ostriker, Kuh & Voytuk, 2011) and to examine the reputation and funding of dissertation advisors and the success (in terms of funding and publication records) of their advisees in more detail. Citation counts for dissertations, user ratings and altmetrics data, e.g., social media data, are valuable indicators of impact that we would like to explore. We also think that productivity and usage datasets can be leveraged to study the emergence of new disciplines and cross-disciplinary subject areas (Sugimoto, Li, Russell, Finlay, & Ding, 2011).

## Acknowledgments

This work was partially funded by the National Institutes of Health under awards NIA P01AG039347 and U01 GM098959. The authors would like to thank and acknowledge the assistance of Samuel Mills in preparing graphics for this text, Mike Gallant for information technology support as well as the ProQuest dissertations product management, development, and technical teams for their support during this research work.

## References

Mazloumian, Amin, Dirk Helbing, Sergi Lozano, Robert Light, and Katy Börner. (2013). "Global Multi-Level Analysis of the 'Scientific Food Web'." Scientific Reports 3, 1167; DOI:10.1038/srep01167.

Ni, C., & Sugimoto, C.R. (2012). Using doctoral dissertations for a new understanding of disciplinarity and interdisciplinarity. Proceedings of the Annual Meeting of the American Society for Information Science and Technology. Baltimore, MD. October 26-30, 2012: ASIST.

Ostriker, J., Kuh, C., & J. Voytuk (Eds.), (2011) A Data-Based Assessment of Research-Doctorate Programs in the United States. Retrieved from: http://www.nap.edu/rdp/

Shneiderman, B. (1992). Tree visualization with tree-maps: 2-d space-filling approach. ACM Trans. Graph. 11, 1 (pp. 92-99). Retrieved from http://doi.acm.org/10.1145/102377.115768

Srivastava, J., Cooley, R., Deshpande, M, & Tan, P., (2000). Web usage mining: discovery and applications of usage patterns from Web data. ACM SIGKDD Explorations Newsletter, 1(2), 12-23. http://doi.acm.org/10.1145/846183.846188

Sugimoto, C. R., Li, D., Russell, T. G., Finlay, S. C., & Ding, Y. (2011). The shifting sands of disciplinary development: Analyzing North American library and information science dissertations using Latent Dirichlet Allocation. Journal of the American Society for Information Science and Technology, 62 (1), 185-204. http://onlinelibrary.wiley.com/doi/10.1002/asi.21435/abstract