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.

Interested to address the diverse needs of university leaders, ProQuest and Indiana University analysed the 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 et al., 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.

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.

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.

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.

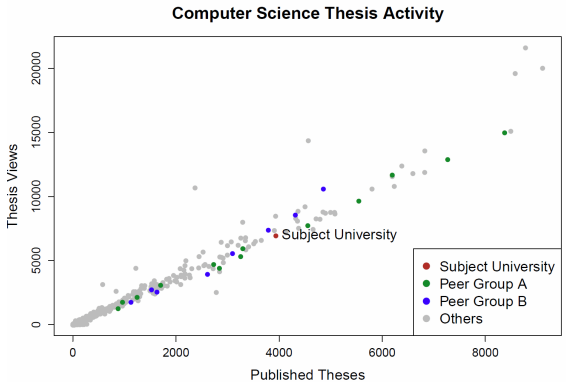

Figure 1. Comparing Subject-Area Specific Thesis Access Activity with Peer Groups.

Study 2: 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.

In Figure 2, 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.

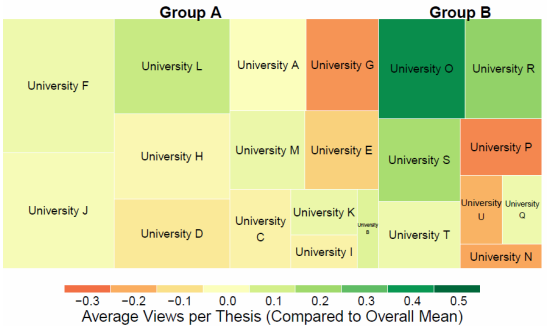

Figure 2. Treemap Comparing Thesis Production and Usage in Computer Science.

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.

Study 3: 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 need to compare this in-flow and out-flow of information with the flows calculated for other universities.

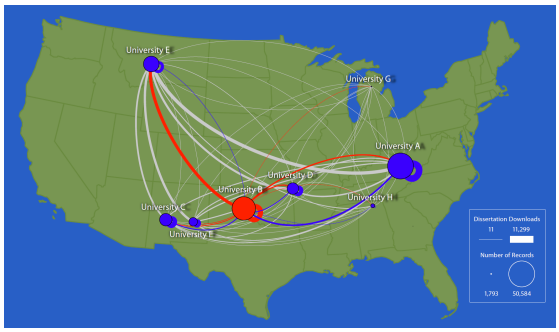

Figure 3. Information Flows within Peer Group

The example in Figure 3 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).

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 P01AG039347, U01GM098959, and U01CA198934. 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, A., Helbing, D., Lozano, S., Light, R. P., & Börner, K. (2013). Global multi-level analysis of the 'Scientific Food Web'. *Scientific reports*, 3.

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