STATEMENT OF SIGNIFICANCE

The objective of this study is to create tools to allow for analysis of topic lifecycles across heterogeneous corpora. While the growth of large-scale datasets has enabled examination *within* scientific datasets, there is a lack of research that looks *across* datasets—examining how different scientific activities enable or propel scientific discovery. Therefore, this work will analyze the importance of various scholarly activities for creating, sustaining, and propelling new knowledge; compare and triangulate the results of three methods for analyzing topics, providing a set of best practices for researchers; and develop transparent and accessible tools that will be made openly available. For years, our knowledge of the scholarly landscape, and subsequently, our understanding of innovation, productivity, and impact, has been largely informed by homogeneous corpora from a single source. However, the growth of datasets that reflect unique areas of scholarly activity have altered the research landscape and provide us with the opportunity to create more accurate understandings of the nature of science. This work should identify which scholarly activities are most indicative of emerging areas and, thereby, identify datasets that should no longer be marginalized, but built into our understandings and measurements of scholarship.

This project will examine the development of topics in four domains (history of science, social network analysis, cognitive science, and digital humanities). These domains represent established areas of research that have combined people, methods, and theories from multiple domains. Examination of the lifecycle of topics in these domains should provide insights into how scholarship evolves across genres in the social sciences and humanities. Triangulation of the methods (word analysis, topic modeling, burst detection, and survival analysis) will be used to ensure the highest level of validity. This project is innovative in its combination of datasets; this research will combine data from formal sources (dissertations, conference proceedings, journal articles, and grant proposals) and informal communication channels (listservs, blogs, twitter, etc.) in order to provide a more holistic lens on scientific communication. Previous investigations have focused on a single source and have marginalized disciplines and researchers who communicate in other ways.

The team is well-qualified to undertake this research, having significant experience with the methods and datasets to be used. The partnership will build upon existing ties; however, the funding of this proposal will likely strengthen and extend the partnership beyond the grant. Cassidy Sugimoto will lead the USA team, with Staša Milojević and Ying Ding. Vincent Lariviere will direct the Canada team and Mike Thelwall will manage the UK team. Each team has strong institutional support in terms of technical and human resources that will be devoted to the project. In particular, all teams will heavily incorporate students into the research, with the funding of the Student Emissaries program. Student Emissaries will not only work with the PIs, but will travel to meetings and have a student retreat that will allow them to share their own research and engage in peer mentoring. At the heart of this initiative is the belief that the best investment we can make in the future of the scientific enterprise is in our students. If science is to become globalized, we must train the next generation in the practice of international collaboration.

The results of this work will have implications for policy makers, as they seek to identify emergent areas of research. It will also provide an indicator of the importance of certain communication channels for identifying emerging areas of knowledge—this provides an array of potential data sources that may serve as leading indicators of topic development. In addition, this project will be useful for scholars; successful researchers need to be good at predicting trends in research and differentiating between topics and research questions that are appropriate for grants, PhD students, and research projects.

The values of the open access movement will be incorporated into the creation of datasets, tools, and scholarly products from this work. Where possible, products will be made available online and we will seek to incorporate the voice of the academic community at all points in our research. We will also work to disseminate the products of the project to a wide community—seeking specialized and general venues for dissemination, favoring those with gold open access policies.