Journal of Informetrics 14 (2020) 101018

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Application of machine learning techniques to assess the trends and alignment of the funded research output

Ashkan Ebadi[a,b,*], Stéphane Tremblay[a], Cyril Goutte[a], Andrea Schiffauerova[b]

[a] National Research Council Canada, Ottawa, Ontario K1K 2E1 Canada
[b] Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montréal, Québec H3G 2W1 Canada

ARTICLE INFO

Article history:
Received 21 May 2018
Received in revised form 8 January 2020
Accepted 20 January 2020
Available online 7 February 2020

Keywords:
Text mining
Topic modeling
Machine learning
Funded research
Publications
Government research priorities
Canada

ABSTRACT

Research and development activities are regarded as one of the most influencing factors of the future of a country. Large investments in research can yield a tremendous outcome in terms of a country's overall wealth and strength. However, public financial resources of countries are often limited which calls for a wise and targeted investment. Scientific publications are considered as one of the main outputs of research investment. Although the general trend of scientific publications is increasing, a detailed analysis is required to monitor the research trends and assess whether they are in line with the top research priorities of the country. Such focused monitoring can shed light on scientific activities evolution as well as the formation of new research areas, thus helping governments to adjust priorities, if required. But monitoring the output of the funded research manually is not only very expensive and difficult, it is also subjective. Using structural topic models, in this paper we evaluated the trends in academic research performed by federally funded Canadian researchers during the time-frame of 2000–2018, covering more than 140,000 research publications. The proposed approach makes it possible to objectively and systematically monitor research projects, or any other set of documents related to research activities such as funding proposals, at large-scale. Our results confirm the accordance between the performed federally funded research projects and the top research priorities of Canada.

1. Introduction

After the Second World War, the developed countries started fully realizing the advantage of research while devoting more and more money to research and development (R&D) activities. Nowadays, a huge amount of money is being annually invested in R&D with the aim of fostering scientific development. Scientists publish their results in the form of scientific papers to secure their priority in discoveries (De Bellis, 2009), in return for the research funding that they have received. With the recent advancements in information technology (IT), the scope of digital data has expanded tremendously. Considering the growing trend of scientific publications (Ebadi & Schiffauerova, 2016a) and the more complex nature of modern science which has become highly interdisciplinary, scientific evaluation calls for complex algorithms able to handle large-scale data. Intensive collaboration, worldwide growth of the number of researchers, and fierce competition for securing the limited financial resources are some of the factors that have given rise to a critical need for publishing at any cost. Hence, apart from the importance of systematic performance evaluation of the funded researchers analyzing the research projects and

* Corresponding author.
E-mail addresses: ashkan.ebadi@nrc-cnrc.gc.ca, a_ebad@encs.concordia.ca (A. Ebadi).

https://doi.org/10.1016/j.joi.2020.101018
1751-1577/

their trends would be necessary for the funding agencies to assure that the projects are in accordance with their visions and strategies.

Quantitative analysis of publication data has been widely carried out in the scientometrics field, revealing important relationships among various factors related to authors and articles. Although the exponential growth of data in the digital era has provided a unique opportunity for research analysts, the complex and highly dimensional nature of the data has made the analyses so challenging such that it rendered the traditional methods unsuitable. However, computer science algorithms provide researchers with novel opportunities to explore new directions for information science. One of the machine learning techniques that can be applied for this purpose is called clustering. Clustering, that is widely in use in various scientific fields and applications (*e.g.* bioinformatics, image segmentation, and document summarization), is an unsupervised learning technique that discovers the hidden groupings in a dataset (Hartigan, 1975). Topic modeling, in general, can be regarded as a clustering technique in which a collection of documents is automatically organized into a set of clusters, so-called topics (Papadimitriou, Raghavan, Tamaki, & Vempala, 2000). Clustering a large volume of text data has several unique challenges compared to non-text data mining tasks. The two main concerns are: 1) the highly unstructured nature of the text that requires encoding to be converted to a format that is recognizable by the clustering techniques, and 2) the high dimensionality of the encoded data (Millar, Peterson, & Mendenhall, 2009).

Topic modeling leverages the context to infer hidden patterns in the unstructured text data, thus providing us with a better understanding of the content and themes of publications. Topic modeling has been widely used in several domains for automatic extraction of semantic or thematic topics from a large corpus of documents (*e.g.* Ebadi & Schiffauerova, 2016a, 2016b; Weng et al., 2010). Additionally, topic modeling is a flexible technique and can be easily generalized to other data types in different domains, *e.g.* image analysis, survey data (Erosheva, 2002), and biological data (Pritchard, Stephens, & Donnelly, 2000). The extracted topics reveal hidden themes in the examined data and can be used for categorizing documents in a large corpus or analyzing the trends of the research projects in a given area. For example, Blei and Lafferty (2006) developed probabilistic time series models to analyze the evolution of topics in the *Science* journal from 1880 to 2000. In another study, Griffiths and Steyvers (2004) focused on the abstracts of articles published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS) and presented a statistical method to discover and assign a set of topics to each document in the corpus. In a more recent study, Weng et al. (2010) analyzed the influential users of Twitter and assessed the topical similarity between users as well as the link structures. They proposed a PageRank-like algorithm for measuring the influence of the Twitterers.

Inspecting the evolution of research topics and their trends could be beneficial for both researchers and policymakers. Topic is one of the first things in a paper that attracts the attention of the reader. However, publications may cover more than one topic, and although it is possible for scientific experts to be aware of the hot topics in their field, due to the large volume of data it is almost impossible for a researcher to know all the interesting topics in all the related scientific fields, or even in a specific field. Summarizing the overall collection of scientific publications in a limited number of themes makes it easier for researchers to get oriented within the areas of new scientific advancements. Additionally, analyzing the research trends enables policymakers to gain a concise picture of projects at the macro level assisting them in directing and adjusting the strategies towards the research priorities.

1.1. Background and related works

Public funding organizations establish strategic roadmaps to support their research and development policies (Gal, Thijs, Glänzel, & Sipido, 2019). Current funding policies target multiple aspects of research activities such as promoting innovation (European Commission, 2017), fostering collaboration, and maintaining and/or improving the country's position in strategic areas (Government of Canada, 2017). For example, the development of emerging technologies that are of high risk and need long-term support could be more dependent on governmental investment (Paull, Wolfe, Hébert, & Sinkula, 2003).

In Canada, three main federal funding organizations support research. The Natural Sciences and Engineering Research Council (NSERC), established in 1978, offers a wide range of programs supporting academic researchers, active in natural sciences and engineering, and fosters innovation through academia-industry linkage. The Social Sciences and Humanities Research Council of Canada (SSHRC), formed in 1977, is the governmental body in supporting social sciences researchers. The Canadian Institute of Health Research (CIHR), established in 2000, is responsible for funding health and medical research. These three agencies together are referred to as the *Tri-Council* (Brook & McLachlan, 2008).

There exist several studies that focused on a specific research domain, discovered topics and performed trend analysis (Chen & Zhao, 2015; Shin, Choi, & Lee, 2015; Zhang et al., 2016). For example, using topic models, Gatti, Brooks, and Nurre (2015) analyzed 80,757 scientific publications' abstracts in the field of operations research and management science, and identified journal groups with similar content. Their analysis also revealed how journals have changed over time, in terms of scope and the topics covered, indicating a significant temporal dynamic. In another study, Park and Song (2013) applied topic models to study research trends in the field of library and information science in Korea within 1970 and 2012. They found an increasing trend of topics such as meta-data and the internet. Sun and Yin (2017) analyzed transportation research trends using topic modeling. They extracted 50 topics from the publications' abstracts published in 22 transportation journals. Their analysis discovered the existence of a few general topics with increasing popularity over time. Yang, Chang, and Choi (2018) collected about 2900 smart factory publications and compared the trends in Korea with international research trends. They

suggest their results could confirm the feasibility of using topic modeling techniques to extract and compare research trends at scale.

The breadth of funded research at the national level is astounding and the number of generated publications is massive. The manual assessment, monitoring, and coding the funded research output is therefore challenging, if not impossible. Despite some recent efforts in leveraging computer science algorithms to design intelligent and automatic research evaluation platforms (Ebadi & Schiffauerova, 2016b), automatic evaluation of the funded research projects trends in the entire domain of natural sciences and engineering, as well as their evolution, at the country level has not yet been accomplished. Moreover, although there exist several studies that investigated research funding systems (Kulczycki, Korzeń, & Korytkowski, 2017; van den Besselaar, Heyman, & Sandström, 2017), relationship between federal funding and research output (Ebadi & Schiffauerova, 2016a; Godin, 2003), or collaboration (Clark & Llorens, 2012; Ubfal & Maffioli, 2011), to the best of our knowledge there is no study that relates national top research priorities with the funded research projects. This important link is missing in similar studies, limiting the scope to only extracting research topics or analyzing their trends. Our proposed approach would enable policymakers to maintain, modify, adjust, or set new strategies in order to achieve their strategic research goals. Lastly, we found no similar study that employed topic modeling, with proper human intervention and hyper-parameter tuning, on large-scale country-level publication data.

In this paper, drawing on large-scale computational data and methods, we focused on the Canadian researchers active in natural sciences and engineering, within the period of 2000–2018. Using structural topic models (STM) we automatically extracted the research themes of the federally funded Canadian researchers' publications and analyzed their trend and evolution. The remainder of the paper proceeds as follows: *Data and Methodology* section describes the data and techniques in more detail; the *Results* section presents the findings of the research; the paper concludes in the *Conclusion* section and some future directions and limitation of the research are discussed in the *Limitations and Future Work*.

2. Data and methodology

The scope of this research covers all Canadian researchers in natural sciences and engineering. Since we intended to study the accordance of the federally funded research topics with the research policies of the country, we focused only on those researchers who were funded by the main federal funding agency in Canada, *i.e.* NSERC, from 2000 to 2018. This sheds light on the alignment of the country's policies and research projects in the 21st century. Based on the data provided by NSERC, we created a database of all the researchers and their federally funded projects. Then we collected publications that acknowledged NSERC as a source of funding from the Elsevier's Scopus which is the largest abstract and citation database of peer-reviewed publications. This data contained all the metadata on the publications such as article id, title, abstract, year of publications, and coauthors. We filtered out publications with no abstract available. Several preprocessing steps were taken on the collected data, *e.g.* correcting special characters and parsing affiliations. Moreover, we performed author disambiguation described by Ebadi and Schiffauerova (2015a) to link the funding and publications datasets. For this purpose, we used full author names and affiliations provided by the NSERC funding data as well as current and past affiliations of authors provided by Scopus. Next, a similarity measure was defined based on author names, affiliations, and research areas. The measure was then used by a machine learning algorithm to disambiguate authors and link the funding and publication datasets. The final database contains 140,966 publications.

We then merged publications titles and abstracts. The integration of abstracts and titles provided us with more information since abstract is a condensed representation of articles and contains more information compared to publication titles, however, titles may also contain some specific keywords or key phrases about the research. Next, we applied several text processing steps to make the textual data appropriate for the algorithm. In particular, we transformed the text to lower case, removed stop words using a customized English stop words list. Numbers, html tags, punctuation marks, and words with less than three characters were all eliminated. Finally, we tokenized the textual data and created a document-term frequency matrix. Having the text data transformed, we performed topic modeling to extract the main research themes.

The conventional methods, such as content analysis (Krippendorff, 2018), may require intensive manual efforts that make them not well-suited for large-scale country-wide analysis (Gatti et al., 2015). On the other hand, machine learning techniques are able to summarize the corpus automatically and extract knowledge from textual data (Blei, Ng, & Jordan, 2003). Topic modeling is an unsupervised machine learning algorithm able to find latent semantic topics in huge and unstructured collections of text data. The model learns the hidden topics through clustering the words with a similar context (Landauer, McNamara, Dennis, & Kintsch, 2013) to soft-cluster the data. The unsupervised learning approach is best applied when although we are interested in the contents of a large corpus, we do not have precise expectations about the structure of texts in the corpus (Lucas et al., 2015). Therefore, it is necessary to choose the algorithm and set the number of topics to perform topic modeling, although there is no correct specification in the general sense (Lucas et al., 2015), and such decisions/choices are highly dependent on the research project objectives.

A family of topic models has recently evolved from Latent Dirichlet Allocation (LDA), a generative Bayesian probabilistic model introduced by Blei et al. (2003) that learns a predefined number of latent topics. We used structural topic models (STM) as some of its properties over LDA were critical for our research objectives. In STM, topics can be correlated and document-level covariates of interest are used to incorporate meta-data (*e.g.* year of publication, in our case) into the topic inference procedure (Roberts, Stewart, & Tingley, 2014). These properties enabled us to capture the temporal aspect hidden in our data and analyze topic distribution trends as well as research projects evolution. The fact that STM allows incorporating

information about each document directly into the topic model enables us to test the hypothesis if, for example, a certain research topic is more likely to be focused later on in time, while another topic is more likely to be observed early on (Lucas et al., 2015). The STM builds a transparent and replicable model relatively fast that requires minor assumptions about the corpus under study (Roberts, Stewart, Tingley et al., 2014). Additionally, STM also takes the correlation between topics into the account. This distinguishes STM from many other mixed membership models. The correlation between topics provides further insights into the structure of the topics at the corpus level (Lucas et al., 2015). In the resulted model each publication is a mixture of topics and the prevalence of the topics are varied according to the incorporated covariates. Similar to other topic modeling approaches and as an unsupervised technique, no data labeling is required in STM and the topics emerge automatically.

The STM has recently attracted the attention of the text mining community. It has been used in various domains, applications, and projects (Bagozzi & Berliner, 2018; Chandelier, Steuckardt, Mathevet, Diwersy, & Gimenez, 2018; Clare & Hickey, 2019; Grajzl & Murrell, 2019; Kuhn, 2018), and it is argued as a high-performing model. Built on the standard LDA (Blei et al., 2003), the structural topic modeling combines and extends the Correlated Topic Models (Blei & Lafferty, 2007), the Dirichlet-Multinomial Regression (Mimno & McCallum, 2012), and the Sparse Additive Generative (Eisenstein, Ahmed, & Xing, 2011) topic models. The topic proportions within documents are controlled by a random variable drawn from a log-normal distribution at the document level, therefore the distribution is not common across all documents (Roberts, Stewart, Tingley et al., 2014). Additionally, word proportions within a topic are not the same across the corpus in STM, and a multinomial logit model is used, thus, a word prevalence is determined by topics, covariates, and topic-covariate interactions (Kuhn, 2018).

One of the key challenges in topic modeling is setting the number of topics, as it is a fixed parameter and should be provided to the algorithm. Fine-tuning this parameter was not considered in several studies (*e.g.* Savoy, 2013; Sugimoto, Li, Russell, Finlay, & Ding, 2011; Yan, 2014), although the number of topics affects the model and the results accordingly (Roberts, Stewart, Tingley, 2014, Arun, Suresh, Veni Madhavan, & Narasimha Murthy, 2010). There are several common approaches in the literature for identifying the optimal number of topics (*e.g.* Arun et al., 2010; Cao, Xia, Li, Zhang, & Tang, 2009; Griffiths & Steyvers, 2004). An example of the common approaches is testing different numbers of topics and using an estimation method such as maximum likelihood to identify the best value for the number of topics (Griffiths & Steyvers, 2004). In addition to the diversity of approaches, a complete automatic framework for identifying the optimal number of topics might not be very accurate (Maskeri, Sarkar, & Heafield, 2008), as finding the exact number of topics is not a trivial task (Grimmer & Stewart, 2013) and often needs human expert intervention. In this study, we applied a semi-automatic multi-layer approach. We first used two metrics, as defined in Cao et al. (2009) and Deveaud et al. (2014), to find a range for the number of topics. We will refer to these metrics as *C-measure*, and *D-measure* in the rest of the paper. The C-measure is calculated through adaptively selecting the best topic model based on density. The intuition is to perform several iterations varying the number of topics and calculating the similarities between topic pairs over several instances of the topic model. The optimal number of topics is then obtained if the similarity between topics achieves its minimum value. The D-measure calculation is slightly different. It uses a heuristic to estimate the number of topics by maximizing the information divergence between all the topic pairs. The optimal values of the C- and D-measure determined the range for number of topics. For example, if 6 and 12 are indicated as the optimal number of topics for the corpus by the C- and D-measure respectively, we considered [6, 12] as the optimal range. If both C and D measures obtained the same value, we defined the range as [value-1, value+1]. In none of our cases the C and D measures differed significantly and we found [6, 12] to be the optimal range based on the mentioned metrics. Next, we followed the data-driven approach proposed by Roberts, Stewart, Tingley (2014) to further refine the range. We ran several automated tests and compared the results using multiple criteria such as held out likelihood (Wallach, Murray, Salakhutdinov, & Mimno, 2009), residuals (Taddy, 2012), semantic coherence, and exclusivity of the topics. In addition, we checked top words and phrases assigned to each topic, topic-document distributions, as well as topic-word distributions to further evaluate the quality of the topics. Based on the performed analysis, it was observed that the best number of topics for our examined corpus is either 8 or 10. Finally, we verified the models with 8 and 10 number of topics by three domain experts that concluded the best number of topics for our corpus, considering our research objectives, is 8. Having set the number of topics, we employed Mimno's coherence measure (Mimno, Wallach, Talley, Leenders, & McCallum, 2011) to further investigate the coherence of the extracted topics. Basically, Mimno's coherence measure calculates the probability of a "rare word" given a "common word" and examines how likely the keywords of a topic occur together. Values closer to zero represent higher coherence.

The STM technique does not assign a representative label to the extracted topics. Although assigning a label to a set of keywords could be relatively easy for a domain expert, finding meaningful and concrete labels for the discovered topics automatically is very challenging (Mei, Shen, & Zhai, 2007). On the other hand, assigning labels manually is subjective (Lau, Grieser, Newman, & Baldwin, 2011). There have been many attempts in the literature to generate labels automatically for topic models (Herzog, John, & Mikhaylov, 2018; Lau et al., 2011; Magatti, Calegari, Ciucci, & Stella, 2009; Mehdad, Carenini, Ng, & Joty, 2013; Mei et al., 2007). Most of the approaches follow two main steps: 1) generating candidate labels often based on measures such as top-n topic words (Lau, Newman, Karimi, & Baldwin, 2010) or the most frequent phrases or n-grams (Mei et al., 2007), and 2) ranking the generated labels (Hulpus, Hayes, Karnstedt, & Greene, 2013; Mei et al., 2007). Although such approaches find a label for a given topic objectively, they are mostly based on the assumption that the top topic keywords are coherent (Lau et al., 2010). But in reality, the top words for the topics are not always coherent. Additionally, in our study, we were interested in finding the main research topics, *i.e.* more generic and inclusive topics, preferably with a simple and interpretable name rather than a phrase. This requirement could not be satisfied with automatic labeling approaches.

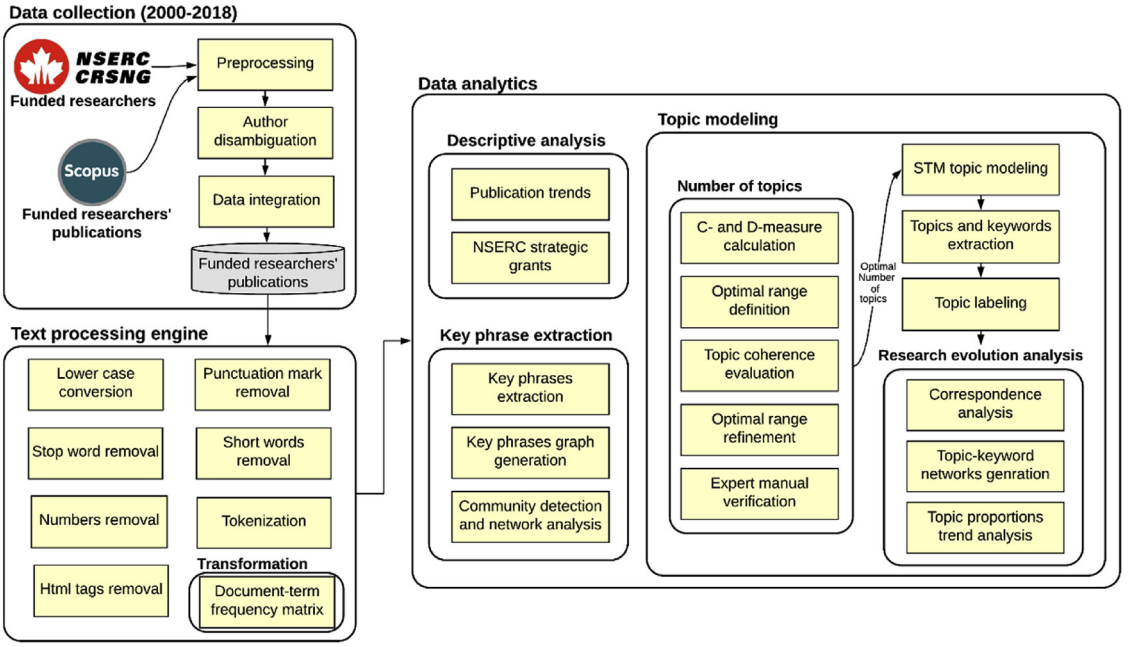

Fig. 1. The analytical flow. The pipeline contains three main components, *i.e.* data collection, text preprocessing engine, and data analytics. In the data collection component, the funded researchers' publications within the period of 2000–2018 are collected. Abstracts and titles of the publications are then merged and preprocessed to make it ready for topic modeling. After performing descriptive analytics and identifying key phrases in the corpus, the optimal number of topics is determined and set in the final STM model to extract the topics and keyword sets. In the final component, the trends and evolution of the research topics are analyzed.

Therefore, we asked three domain experts to manually examine the extensive set of keywords for each topic and assign a representative label to the topic. First, the experts separately analyzed the list of keywords and identified the topic labels. Next, all three experts shared their labels and tried to convince other researchers if there was an incompatibility. Finally, the label with the highest vote on a topic was chosen as the topic's label. We used more than one expert to reduce the subjectivity effect.

Before analyzing the prevalence of discovered research topics over time, we analyzed the key phrases for each year of the examined time interval, using social network analysis, to ensure the existence of temporal trends in publications. We then confirmed that by applying correspondence analysis on the keywords. The analytic pipeline is shown in Fig. 1. The entire pipeline was coded in Python and R programming languages. Gephi software (Bastian, Heymann, & Jacomy, 2009) was used to create topic-key phrases graph visualization.

3. Results

3.1. Descriptive analysis

As the main federal funding agency in Canada, NSERC funded more than 97,000 distinct researchers within the period of 2000–2018. We extracted 140,966 publications of the funded researchers, listed in the Scopus database.

3.1.1. Funded researchers' publications

Fig. 2 shows the distribution of extracted articles over the examined period. The bolded numbers on the bars show the total number of publications. We tagged an article as a Canadian publication if at least one of the authors had a Canadian affiliation. Canadian and international publications were color-coded in red and blue in Fig. 2, respectively. As seen, the number of publications follows an increasing trend, exceeding more than five-fold comparing the beginning and end of the examined period. This is in line with several studies that confirm the existence of an increasing scientific publication trend (*e.g.* Bornmann & Mutz, 2015; Zeng et al., 2019). Although the number of publications is almost constant in the early years, a sudden increase is observed in the final years. This drastic increase might be partially due to the higher coverage of the publications in Scopus in recent years as well.

3.1.2. NSERC strategic grants

The Strategic Partnership Grants (SPG) of NSERC is one of the funds that focuses on targeted areas and is highly in line with the country's macro research policies (NSERC, 2017b). The program also aims to foster research collaboration between academia, industry, and government. The strategic grants have four main target areas: 1) advanced manufacturing,

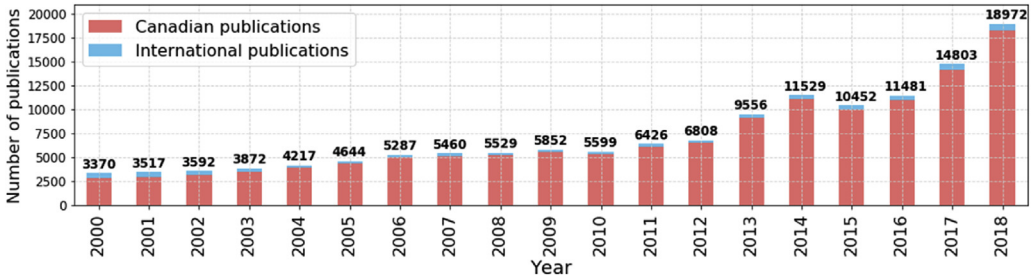

Fig. 2. The trend of the funded researchers' publications between 2000 and 2018. Bolded values on the bars indicate the total number of publications. Red and blue bars represent Canadian and international publications, respectively (for interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

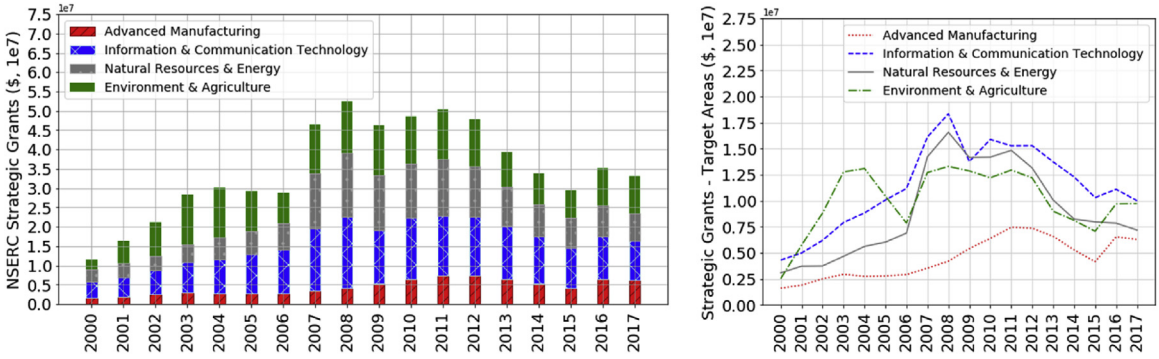

Fig. 3. NSERC strategic projects grants by main target areas in the 2000–2017 period.

2) environment and agriculture, 3) information and communication technologies, and 4) natural resources and energy. Fig. 3 depicts the trend of the mentioned categories within the period of 2000–2017. In the figure, years represent fiscal years, *e.g.* 2017 represents the fiscal year of 2017-2018. As seen in Fig. 3a, NSERC strategic grants followed a sinusoidal trend over the examined period. The lowest proportion was allocated to advanced manufacturing over the years. According to Fig. 3b, the advanced manufacturing funding has less fluctuated compared with the other three target areas, following an almost constant trend over the beginning years. However, the other categories have experienced more fluctuations, peaked in 2008 and after that followed a decreasing trend. The slope of the increasing trend for all the categories became steeper around 2006 and the declining trend started around 2011. The information and communication technology (ICT) research has received the most proportion of funding, except for the beginning 5 years where the environment and agriculture were the top priority. Interestingly, the funding trend for ICT and environmental projects reached almost the same level in 2017.

3.2. The generated topic model

As explained in the "Data and Methodology" section, we followed a multi-layer approach to find the number of topics for our corpus and found the optimal number of topics is 8. We then applied structural topic models to extract the main scientific areas that the funded researchers have focused on from 2000 to 2018.

3.2.1. Topic coherence

Topic coherence evaluation is one of the aspects to assess the quality of the topics discovered by topic models (Rosner, Hinneburg, Röder, Nettling, & Both, 2014). We used Mimno's coherence measure (Mimno et al., 2011) to evaluate the coherence of the extracted topics. As seen in Fig. 4, the topic coherence measure is very close to zero representing high coherence in the extracted topics. This partially validates the identified number of topics, as if the number of topics was not identified appropriately the coherence measure would be much higher. Based on the calculated measure, topics 3, 4, 5, and 8 are slightly more coherent.

3.2.2. The extracted topics

We extracted eight research topics and used the year of publication as the covariate. Using Bischof and Airoldi (2012) approach, we extracted topic keywords that were both frequent and exclusive. This filtration helps to improve the quality of the keywords as if we only focus on the frequency, frequent keywords are often those words that are needed to discuss any topic, and hence, they will not be very informative and specific. On the other side, only exclusive words might be very rare and as a result not very informative as well. We then verified the topics and assigned them a representative label *via* three

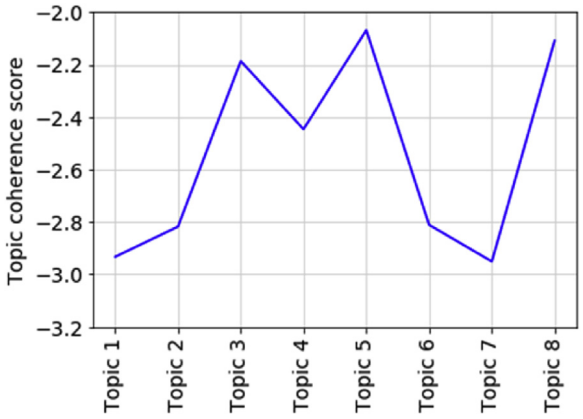

Fig. 4. The coherence of the extracted topics.

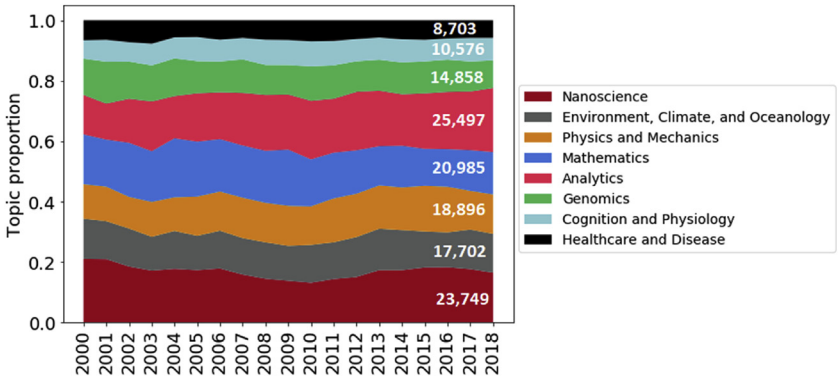

Fig. 5. Dominant topic distribution across publications. The numbers on the figure represent the total number of publications dominated by the respective topic.

domain experts' manual verification of the extensive set of keywords, as defined in the "Data and Methodology" section. The labeled extracted topics are: 1) nanoscience, 2) environment, climate, and oceanology, 3) physics and mechanics, 4) mathematics, 5) analytics, 6) genomics, 7) cognition and physiology, and 8) healthcare and disease. One should note that these eight research topics may only represent the main research areas at an abstract level.

3.2.3. Dominant topic distribution across publications

In STM topic modeling, each document may belong to more than one topic as a topic is assigned to each document with a probability. One of the outputs of the topic model is the publication-topic probability matrix that contains probabilities of observing each topic in each publication. Thus, our publication-topic probability matrix was a p by k matrix where p is the total number of publications and k is the number of topics. We assigned the topic with the highest probability to each document and analyzed the dominant topic distribution across publications. For example, if for publication P_1, the topic probabilities for topics 1 to 8, were (0.1, 0, 0, 0.4, 0.3, 0, 0, 0.2), we assigned topic-4 to P_1. As seen in Fig. 5, analytics, nanoscience, and mathematics were the top-3 most dominating topics. It is also observed that more publications were dominated by analytics along time while an opposite trend is observed for nanoscience and mathematics. Despite some fluctuations, an almost steady trend is seen for the other five topics, namely environmental studies, physics, genomics, physiology, and healthcare.

3.3. Research evolution and prevalence analysis

To investigate the evolution and prevalence of the research topics, we first mapped publications to years using extracted key phrases and applied social network analysis to assess the existence of trends in publications. We then confirmed the trends by applying the correspondence analysis on the keywords. Finally, we analyzed the trends of the publications.

3.3.1. Key phrase analysis

We extracted n-grams of length 2 to 4 for each year of the examined period and created a 2-mode (bipartite) network in which each year is linked to the representative top-30 phrases. We then employed the Louvain modularity method (Blondel,

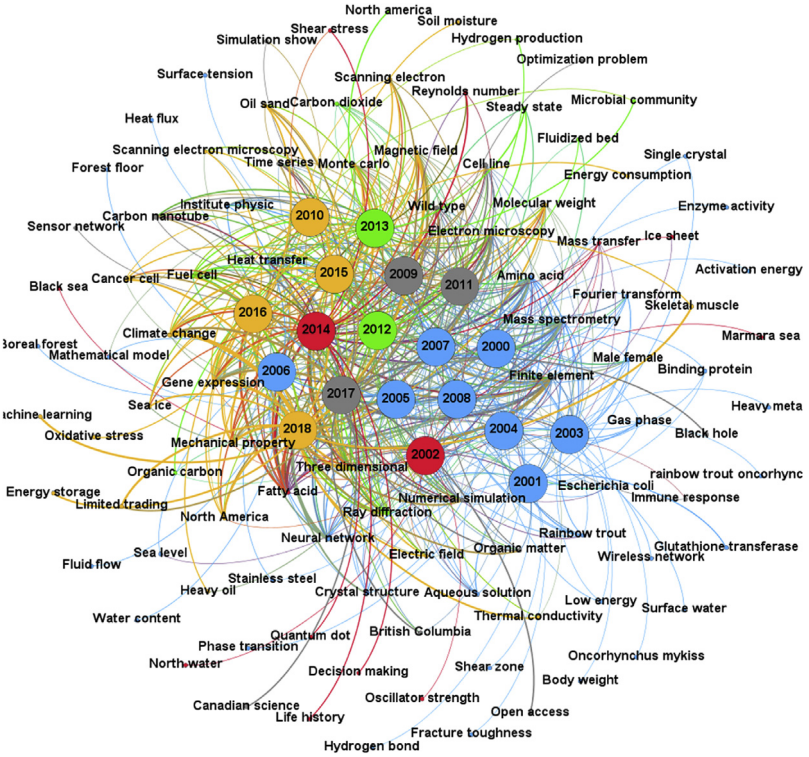

Fig. 6. The < year, key phrase > network for the top-30 n-grams of length 2 to 4. Color represents the network community.

Guillaume, Lambiotte, & Lefebvre, 2008) to detect the community structure of the generated network. The result is depicted in Fig. 6. Colors in the figure represent different communities. The community structure partially implies the existence of annual trends in scientific publications as we see 5 different groups (communities) of years based on the top key phrases. Therefore, there is a relationship between the top key phrases and the year of publication. We will further investigate the existence of such a relationship in the next section.

Additionally, several observations are made on the trend of key phrases. The "*Neural networks*" phrase has been among the top phrases in almost the entire examined period, except for 2013, 2014, and 2015, and was among the top-3 in the beginning four years. The "*Machine learning*" phrase appeared in 2018. Such phrases could partially demonstrate the importance of artificial intelligence and advanced analytics as one of the strategic research areas of Canada. Another example of interesting key phrases is the "*Oil sand*" that appeared in 2009, 2013, and the last four years of the examined period. Oil sand is a mixture of sand, water, and a thick black crude oil called bitumen. The majority of Canada's oil reserves, *i.e.* ~96 percent, are located in the oil sands (BP, 2019) that requires advanced and innovative technology to not only develop the oil sands but also improve environmental performance (Canadian Association of Petroleum Producers, 2019). Climate change has recently attracted the attention of scientific community as the negative impacts of climate change are already being observed. The "*Climate change*" phrase first appeared in 2008 and then has been always among the top-30 phrases since 2010, being the top phrase in 2018. This increasing attention can be regarded as a sign of the high priority of the issue in Canada.

3.3.2. Correspondence analysis

To further verify the existence of annual trends in scientific publications, we used correspondence factor analysis (Greenacre, Blasius, & Blasius, 2006). Correspondence analysis is a multidimensional technique, designed for comparing profiles and patterns through providing a graphical representation of cross-tabulations (Doré & Ojasoo, 2001). By mapping the data on visually understandable dimensions, it filters out the noise and analyzes correlations among variables to provide a graphic output in which it is easy to grasp the patterns (Blasius & Greenacre, 1998).

To perform the correspondence analysis, we first created a term-frequency matrix for each year of the examined period. Next, we filtered out terms that had been appeared less than 300 times each year. We also filtered highly frequent terms that had been appeared in more than 60 % of the documents as we found them to be not very informative. We provided the algorithm with the generated term-frequency matrix that mapped the matrix to a two-dimension space by extracting the first two principal components.

Fig. 7 shows that the studied research projects have evolved over time, forming a U-shape curve. Based on the distance of the variables from the axes and the origin, it is clear that the frequent terms have evolved over time which confirms our findings in the previous section.

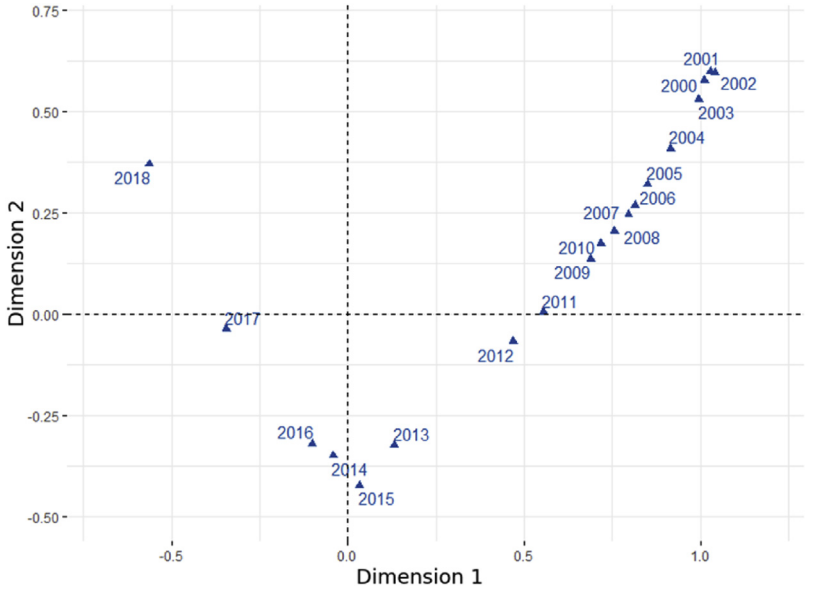

Fig. 7. Correspondence analysis on the frequent terms appeared in the NSERC funded researchers' publications from 2000 to 2018.

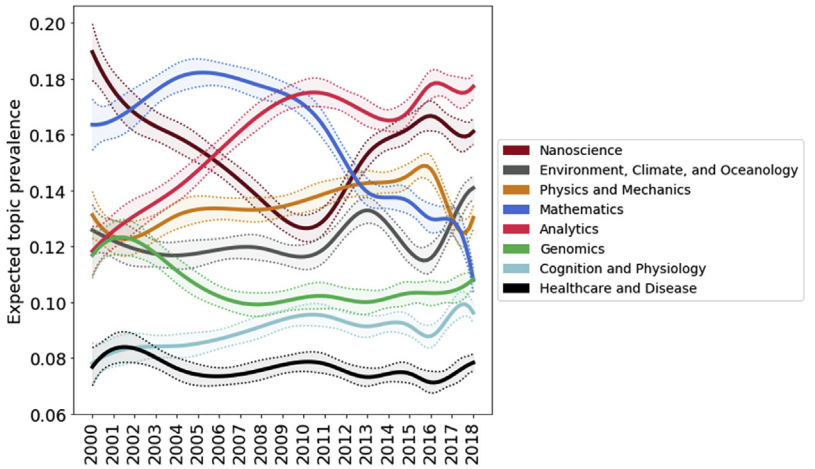

Fig. 8. Topic prevalence from 2000 to 2018. The shaded areas between the dotted lines indicate 95 % confidence interval area.

3.3.3. *Trends of the extracted research topics*

Having verified the existence of an annual trend of NSERC funded researchers in scientific publications, we next analyzed the evolution of the extracted topics, representing the main scientific areas, over time. To evaluate the proportions of topics over time, we used the *estimateEffect* function of the *stm* package in R (Roberts, Stewart, Tingley, 2014). The function regresses the proportion of each publication on a publication-specific covariate (in our case, year of publication). The outcome of the regression model is topic proportions. Using this approach we can estimate conditional expectation of topic prevalence, given the characteristics of publications.

Fig. 8 shows the topic proportions from 2000 to 2018. The shaded areas between the dotted lines in the figure represent the 95 % confidence interval area. It is noticed that although in the beginning, *Nanoscience* was the main focus of NSERC funded researchers' publications, *Analytics* first became the most important topic in 2009 and despite some fluctuations, it maintained the first place afterward. One reason for the decreasing trend of *Mathematics* could be the rise of emerging scientific fields as well as advanced analytics as such fields have, partially or completely, embedded mathematical approaches to achieve the results. Additionally, the more multidisciplinary nature of modern science could also be regarded as another explaining factor for such a decreasing trend. The same justification could be valid for other core scientific fields such as physics. However, the *Physics and Mechanics* topic maintained almost the same level within the examined period, although a recent decline is observed. The trend of *Nanoscience* is quite different from the other topics where it has followed a sinusoidal trend with a peak in 2000 and 2016 and the minimum in 2010. Despite the steady trend in the first decade of the 21st century

and recent fluctuations, environmental issues have again attracted the high attention of NSERC funded researchers, indicated by the proportion trend of the *Environment, Climate, and Oceanology* in the figure. In fact, it was the third research priority in 2018 after *Analytics* and *Nanoscience*. The *Healthcare and Disease* related publications have maintained a steady trend over time, ranking the last in terms of proportion in almost the entire period. After a sudden decrease in 2001, publications on *Genomics* maintained an almost constant trend after 2006. After a slightly increasing trend till 2010, *Cognition and Physiology* topic followed a sinusoidal trend with a drastic increase between 2016 and 2017. Of course, there could be some overlaps between the studies performed in the latter three topics, and even other topics such as *Analytics*.

4. Discussion

Nowadays, digital data at a large scale is highly available. However, manual processing and coding of huge collections of documents such as the scientific output of researchers for funding agencies or any scientific evaluation organization is almost impossible. Even if manual assessment is feasible, it cannot provide a coherent and accurate global view due to the human factor involvement. The highly inter-disciplinary and evolving nature of modern science also adds to the complexity of such an evaluation as small pool of reviewers is available for interdisciplinary research (Lastewka, 2002). Analyzing and monitoring research products using text mining and machine learning approaches could be much cheaper, unbiased, and more informative. Such techniques can not only summarize the main research domains that were studied, they also help to verify if the studied areas were in line with the top research priorities of a country or a funding organization. This can ensure the relevance of the performed research and align them with the government priorities, if required. In Canada, although NSERC is not the only funding body of the government, it is the national instrument for making strategic science and technology investments (Lastewka, 2002). In this paper, we focused on publications of the NSERC funded researchers from 2000 to 2018 and used text mining, topic modeling, and social network analysis to analyze the main research areas as well as their evolution trends.

As one of the world leaders in science, technology, and innovation, Canada generates over 4 % of the global knowledge while its population accounts for only 0.5 % of the world's population (Canadian Trade Commissioner Service, 2018). In the late 1990s, after realizing that the country is losing its scientific competitiveness (Robitaille & Gingras, 1999), Canada started allocating significant funds to scientific activities and employed various strategies such as attracting world-leading scientists to improve the R&D sector (Fast, 2007). In 1997–1998, NSERC invested more than $145 million, about 33 % of its entire budget, to support young researchers and to train the next generation of scientists (Manley, 2000). Such investments were realized in the early years of the 21st century, as funding may take about three to five years to show the impact on the scientific output (Ebadi & Schiffauerova, 2015b). In continuation of the support, NSERC made more than $530 million investment in university-based research and training in 2000–2001 (NSERC, 2001). Collectively, the three main federal funding agencies, *i.e.* NSERC, CIHR, and SSHRC, invested more than one billion dollars only on university-based research and academia-industry collaboration (Lastewka, 2002). Within recent years, NSERC further continued to support university professors, with a special focus on early-career scientists. In 2017–2018, more than 11,700 professors and over 8000 students and postdoctoral fellows were provided with direct support (NSERC, 2018). They have also implemented new policies to encourage multinational and foreign companies' collaboration with Canadian universities. Only in 2017–2018, about 3600 industrial partners participated in NSERC programs, leveraging more than $250 million (NSERC, 2018).

During 2000 and 2010, Canada's Gross Domestic Expenditure on Research and Development (GERD) ratio to the Gross Domestic Product (GDP) was almost stable around 2 %, with a declining trend over the final five years (Findlay & Dodd, 2016). The declining trend continued after 2010, reaching the level of 1.5 % in 2018. This has been Canada's lowest level in the 21st century, in contrast to the highest level of 2.02 % that was allocated in 2001 (OECD, 2018). Although the R&D investments in Canada have diminished, the federal government set policies to support research centers and to foster world-class research programs in universities as well as encouraging industrial R&D investment (Fast, 2007). One may note that natural sciences and engineering accounts for over 90 % of total national expenditure on R&D in Canada (Statistics Canada, 2011).

Within the first decade of the new century, Canada experienced two major strategies on research and innovation activities. In 2001, the federal government released Canada's innovation strategy, setting directions for research and innovation until 2010 with an aim to move toward a more innovative economy (Government of Canada, 2001). Although the general path was well-discussed, there was no specific priority research area highlighted to channel the funding. In other words, there was no clear priority at that time and it was recommended that some high priority research areas need to be established. Therefore, any research that was scientifically solid, without any focus on the needs of the country, had an equal chance to secure funding. Later in September 2006, the Council of Canadian Academies (CCA) released a report, called "*the state of science and technology in Canada*", which had an objective to better focus on research strengths and priorities. According to CCA, environmental issues, natural resources and energy, health and life sciences, and information and communication technologies were identified as the main science and technology priority areas (Council of Canadian Academies, 2006). These areas, as well as other research activities of national importance such as forestry, fisheries, and health, were also supported by different funding programs of NSERC, *e.g.* through the Strategic Project Grants, Strategic Network Grants (NSERC, 2009). In the fiscal year 2000–2001, NSERC allocated 7 % of its expenditures to the strategic projects (Lastewka, 2002). The investment in strategic areas raised steadily in the first decade peaking in 2008, following with a slightly declining trend afterward (NSERC, 2019b). NSERC follows a specific mechanism to choose and modify target research areas of national importance

and considers a five-year cycle to revisit and refine those areas (Lastewka, 2002), by a committee of domain experts (NSERC, 2015).

According to the latest CCA assessment of the state of science, technology, and R&D in Canada (Council of Canadian Academies, 2016), although Canada's international R&D position is strong, the number of R&D personnel in Canadian business side has dropped by 20 % between 2008 and 2013. The decline in business R&D investment could partially explain the steadily declining trend of R&D share of GDP in Canada since 2001 (OECD, 2018). The CCA analysis also confirmed that Canada's research strengths have remained relatively stable since 2012, based on three dimensions, magnitude in terms of share of publications, the impact of the research performed, and growth in the research output (Council of Canadian Academies, 2016). Additionally, they identified cognitive science and healthcare as areas in which Canada is internationally competitive. The report also stated that Canada mostly lags behind other developed countries in research on enabling and strategic technologies with an exception of few interdisciplinary domains in ICT such as health informatics (Council of Canadian Academies, 2016). Even in areas where Canada initially played an important role and retained research capacity over time, *e.g.* artificial intelligence (AI) and regenerative medicine, the country's international research share was not satisfactory (Council of Canadian Academies, 2016).

In 2017 federal budget, Canada set strategies to focus on expanding growth in six priority areas, *i.e.* advanced manufacturing, agri-food, digital industries, clean technology, health/bioscience, and clean resources, to improve the innovative and competitive capacity in science and technology (Canadian Trade Commissioner Service, 2018). Moreover, $125 million (over 5 years) was dedicated to improving the country's position in AI, through the Pan-Canadian AI strategy (Canadian Trade Commissioner Service, 2018). The strategy aims to attract skilled AI researchers, establish a better connection between major AI centers, develop global thought leadership, and support the national AI research community (CIFAR, 2019). A new initiative, named *Superclusters*, was also established to support innovative projects and ground-breaking research through linking industrial companies, academic institutions, and not-for-profit organizations. With an investment of $950 million (over 5 years), matched dollar for dollar by the private sector, superclusters initiative is expected to create over 50,000 jobs over 10 years and grow Canada's GDP (Government of Canada, 2019).

Our findings identified data analytics as the most dominant research area in recent years. We also observed that data analytics has increasingly attracted researchers' attention in terms of the number of publications covering the topic. These findings are highly in-line with the top research priorities of Canada, and NSERC's as well, in supporting AI and advanced analytics. Canada is one of the leaders in ICT research, making a considerable annual investment. Within 2000 and 2011, Canada's nominal ICT investment growth per year outperformed the United States (Capeluck, 2013). In addition to the national strategies to support AI and ICT research, NSERC invested about $200 million in ICT research in 2017–2018, supporting several research topics such as communication networks and services, internet of things, advanced data management and analytics, cybersecurity, human interaction with digital media, and quantum computing. Additionally, 601 industrial partners contributed about $45 million to ICT research in the mentioned period (NSERC, 2019a).

Canada ranks sixth in the world in energy production, having the United States as the main export destination for energy products (Natural Resources Canada, 2019b). Several energy and physics R&D activities are being done in close collaboration between the US and Canada, *e.g.* the clean energy dialogue launched in 2009 (Hale, 2011). As the second largest producer of hydroelectricity in the world (Natural Resources Canada, 2019a), the fact that Canada is a world-leader in energy production was clearly reflected in physics and energy-related topics as well as key phrases of NSERC funded research projects.

Genomics is also one of the research priorities of Canada. Established in 2000, Genome Canada is one of the main organizations working on genomic-based technologies aiming to develop and implement a national strategy for genomic research projects (GenomeCanada, 2015). The organization collaborates with six genome centers in the country, *i.e.* Genome British Columbia, Genome Alberta, Genome Prairie, Ontario Genomics, Génome Québec and Genome Atlantic, leading Canada's genome research. Recent advances in computer science have made this field highly inter-disciplinary involving computer scientists and bioinformaticians working on computer aided diagnosis (CAD) systems using genome profiling data. This resulted in the emergence of new high-tech inter-disciplinary scientific areas such as medical informatics where scientists apply complex informatics and analytics approaches to solve medical problems. These areas are greatly contributing to the improvement in patients' outcomes and medical diagnoses. Through the Collaborative Research and Training Experience (CREATE) program, NSERC has recently targeted the collective potential of the participating universities to train contributors to healthcare research, aiming to improve the healthcare system through facilitating advanced medical informatics (NSERC, 2017a). We found genomics as one of our main extracted topics, with an increasing share in recent years.

By applying text mining, natural language processing, social network analysis, and machine learning techniques on a large volume of data, we proposed a solution for monitoring research activities and checking their alignment with the top research priorities of a country. To the best of our knowledge, this study goes beyond previous literature by linking top national research priorities with the extracted research projects. This critical link was missing in similar studies as their scope was limited to only discovering research topics using topic modeling or analyzing their trends. Our implemented analytics pipeline, which automatically discovers hidden topic-based patterns and extracted them from the collection of documents, is able to handle text datasets at a very large scale. The proposed system can be employed as a decision support tool to automatically monitor research activities, and ensure industrial and socio-economic relevance of the performed research according to the defined research priorities by the government. We used STM technique to discover the research themes. Although researchers are widely using STM especially for social science topic modeling (Foulds, Kumar, & Getoor, 2015), like all models, it has advantages and disadvantages. Tuning model parameters, *e.g.* the number of topics, could be challenging,

however, the parameter space is not correlated with the corpus size, and therefore, the technique could be appropriate for large scale data as well. Moreover, as a probabilistic generative model, STM can assign probability to a new unseen document. The soft clustering nature of STM technique allows multiple topics to assign to documents, rather than a single topic for each document. In addition, STM algorithm allows incorporating covariates. These properties were of our interest in analyzing the trends of funded researchers' projects. Although we used STM, many of the procedures/steps we introduced/followed could be applied to other similar pipelines. For example, the multi-layer approach for finding the number of topics could be applied directly to a broader class of topic models.

5. Conclusion

The explosion of new sources of data in parallel with the recent rapid development of advanced analytics techniques have created new opportunities to analyze large-scale data (Grimmer & Stewart, 2013). We believe there is a great potential for big data analysis in the social sciences and scientometrics community. As one of many machine learning frameworks, topic models offer an innovative and objective approach to measure latent qualities on large datasets (Lucas et al., 2015). We proposed a research monitoring framework incorporating text mining and topic modeling techniques. Of course, such tools cannot substitute for substantive knowledge of human experts, but they can be used as powerful decision support systems to structure humans' effort and augment their capabilities/efficiency to handle the enormous volume of data on research input and output. The STM algorithm, in particular, allowed us to incorporate document-topic proportions into the model and analyze topic prevalence over time. Our findings suggest that in general the projects of Canadian researchers in natural sciences and engineering were in line with the top research priorities of the country as well as NSERC's. The appearance of environmental and climate projects among the important topics was notable, which indicates the importance of environmental issues in the Canadian scientific community. The country's intensive investment in AI and emerging technologies that was reflected as the rise of analytics in the projects is also noticeable. Finally, the existence of a temporal trend in scientific publications, as well as the highly interdisciplinary nature of modern science, would highlight the need for such a comprehensive analysis to be performed frequently.

6. Limitations and future work

We focused on NSERC funded research within the period of 2000–2018. The findings of this research may only shed light on the research agenda of NSERC and Canada at a very high level. A future direction would be taking sub-disciplines/areas of national interest, such as AI, into consideration and analyze them separately. Additionally, other funding agencies could be included to perform a more comprehensive assessment. This would also enable evaluating the impact of collaboration on emerging new research areas. Our proposed methodology could be applied to any country and any research and funding system, with proper tuning based on the target country and/or funding organization. Another future direction would be to apply the proposed approach in different countries and compare the results. We analyzed the prevalence of certain topics over time. Analyzing the existence of topics fusion and/or division phenomenon over time could be also considered as a future research direction. We used the publications' abstracts and titles to perform the topic modeling. Although the entire paper is summarized in the abstract and title, another future direction would be analyzing the entire body of publications. Analyzing the method sections could be of special interest as it may reveal methodological evolution. Finally, an in-depth analysis of the relations between the spent research money and research output would be suggested.

Author contributions

Ashkan Ebadi: Conceived and designed the analysis, Collected the data, Contributed data or analysis tools, Performed the analysis, Wrote the paper.

Stéphane Tremblay: Wrote the paper, Other contribution.

Cyril Goutte: Wrote the paper, Other contribution.

Andrea Schiffauerova: Wrote the paper, Other contribution.

Acknowledgement

The authors would like to offer their special thanks to Dr. Normand Péladeau (Provalis Research) for his valuable comments.

References

Arun, R., Suresh, V., Veni Madhavan, C. E., & Narasimha Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In M. J. Zaki, J. X. Yu, B. Ravindran, & V. Pudi (Eds.), *Advances in knowledge discovery and data mining* (pp. 391–402). Berlin Heidelberg: Springer.

Bagozzi, B. E., & Berliner, D. (2018). The politics of scrutiny in human rights monitoring: Evidence from structural topic models of US state department human rights reports. *Political Science Research and Methods*, *6*(4), 661–677. http://dx.doi.org/10.1017/psrm.2016.44

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open source software for exploring and manipulating networks. *Third International AAAI Conference on Weblogs and Social Media. Presented at the Third International AAAI Conference on Weblogs and Social Media*,. Retrieved from. https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154
Bischof, J. M., & Airoldi, E. M. (2012). Summarizing topical content with word frequency and exclusivity. *Proceedings of the 29th International Conference on International Conference on Machine Learning*, 9–16. Retrieved from. http://dl.acm.org/citation.cfm?id=3042573.3042578
Blasius, J., & Greenacre, M. (1998). *Visualization of categorical data*. Academic Press.
Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, 113–120. http://dx.doi.org/10.1145/1143844.1143859
Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, *1*(1), 17–35. http://dx.doi.org/10.1214/07-AOAS114
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(January), 993–1022.
Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics Theory and Experiment*, *2008*(10), P10008. http://dx.doi.org/10.1088/1742-5468/2008/10/P10008
Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, *66*(11), 2215–2222. http://dx.doi.org/10.1002/asi.23329
BP. (2019). *Statistical review of world energy* Retrieved August 20, 2019, from BP global website:. https://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy.html
Brook, R. K., & McLachlan, S. M. (2008). Trends and prospects for local knowledge in ecological and conservation research and monitoring. *Biodiversity and Conservation*, *17*(14), 3501–3512. http://dx.doi.org/10.1007/s10531-008-9445-x
Canadian Association of Petroleum Producers. (2019). *Oil sands* Retrieved August 20, 2019, from Canadian Association of Petroleum Producers website:. https://www.capp.ca:443/canadian-oil-and-natural-gas/oil-sands
Canadian Trade Commissioner Service. (2018). *Canada's innovation strengths and priorities* Retrieved August 22, 2019, from. https://www.tradecommissioner.gc.ca/innovators-innovateurs/strategies.aspx?lang=eng
Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7), 1775–1781. http://dx.doi.org/10.1016/j.neucom.2008.06.011
Capeluck, I. (2013). *Canada-US ICT investment in 2011: The gap narrows* Retrieved from Centre for the Study of Living Standards website. http://www.csls.ca/notes/Note2013-1.pdf
Chandelier, M., Steuckardt, A., Mathevet, R., Diwersy, S., & Gimenez, O. (2018). Content analysis of newspaper coverage of wolf recolonization in France using structural topic modeling. *Biological Conservation*, *220*, 254–261. http://dx.doi.org/10.1016/j.biocon.2018.01.029
Chen, H., & Zhao, J. L. (2015). *ISTopic: Understanding information systems research through topic models (SSRN scholarly paper No. ID 2601719).* Retrieved from Social Science Research Network website:. https://papers.ssrn.com/abstract=2601719
CIFAR. (2019). *Annual report of the CIFAR Pan-Canadian AI strategy* Retrieved from. https://www.cifar.ca/docs/default-source/ai-reports/ai_annualreport2019_web.pdf?sfvrsn=244ded44_17
Clare, S. M., & Hickey, G. M. (2019). Modelling research topic trends in community forestry. *Small-scale Forestry*, *18*(2), 149–163. http://dx.doi.org/10.1007/s11842-018-9411-8
Clark, B. Y., & Llorens, J. J. (2012). Investments in scientific research: Examining the funding threshold effects on scientific collaboration and variation by academic discipline. *Policy Studies Journal*, *40*(4), 698–729. http://dx.doi.org/10.1111/j.1541-0072.2012.00470.x
Council of Canadian Academies. (2006). *The state of science and technology in Canada* Retrieved August 28, 2019, from. https://cca-reports.ca/reports/the-state-of-science-and-technology-in-canada/
Council of Canadian Academies. (2016). *Competing in a global innovation economy: The current state of R&D in Canada—Expert panel on the state of science and technology and industrial research and development in Canada* Retrieved from. http://new-report.scienceadvice.ca/assets/report/Competing_in_a_Global_Innovation_Economy_FullReport_EN.pdf
De Bellis, N. (2009). *Bibliometrics and citation analysis: From the science citation index to cybermetrics*. Scarecrow Press.
Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, *17*(1), 61–84.
Doré, J.-C., & Ojasoo, T. (2001). How to analyze publication time trends by correspondence factor analysis: Analysis of publications by 48 countries in 18 disciplines over 12 years. *Journal of the American Society for Information Science and Technology*, *52*(9), 763–769. http://dx.doi.org/10.1002/asi.1130
Ebadi, A., & Schiffauerova, A. (2015). How to become an important player in scientific collaboration networks? *Journal of Informetrics*, *9*(4), 809–825. http://dx.doi.org/10.1016/j.joi.2015.08.002
Ebadi, A., & Schiffauerova, A. (2016). How to boost scientific production? A statistical analysis of research funding and other influencing factors. *Scientometrics*, *106*(3), 1093–1116. http://dx.doi.org/10.1007/s11192-015-1825-x
Ebadi, A., & Schiffauerova, A. (2015). How to Receive More Funding for Your Research? Get Connected to the Right People!. *PloS One*, *10*(7), e0133061 http://dx.doi.org/10.1371/journal.pone.0133061
Ebadi, A., & Schiffauerova, A. (2016). iSEER: An intelligent automatic computer system for scientific evaluation of researchers. *Scientometrics*, *107*(2), 477–498. http://dx.doi.org/10.1007/s11192-016-1852-2
Eisenstein, J., Ahmed, A., & Xing, P. (2011). *Sparse additive generative models of text*. http://dx.doi.org/10.1184/R1/6476342.v1
Erosheva, E. A. (2002). *Grade of membership and latent structure models with application to disability survey data* Retrieved from. Department of Statistics, Carnegie Mellon University. https://www.stat.washington.edu/elena/papers/Erosheva-thesis-2002.pdf
European Commission. (2017). *Europe's future: Open innovation, open science, open to the world* Retrieved from Reflections of the Research, Innovation and Science Policy Experts (RISE) website:. https://www.fct.pt/noticias/docs/Europe_s_future_Open_Innovation_Open_Science_Open_to_the_World.pdf
Fast, E. (2007). *Mobilizing science and technology: The new federal strategy Parliamentary Information and Research Service* Retrieved from Library of Parliament (Canada) website:. https://www.sreducation.ca/wp-content/uploads/2012/09/Library-of-Parliament-Federal-Strategy-for-Science-and-Technology.pdf
Findlay, S., & Dodd, M. (2016). *Science and technology in Canada. Retrieved from social sciences and humanities research council website*. http://ifsd.ca/web/default/files/Policy%20Briefs/Policy%20Brief%20-%20Science%20(English).pdf
Foulds, J., Kumar, S., & Getoor, L. (2015). Latent topic networks: A versatile probabilistic programming framework for topic models. *International Conference on Machine Learning*, 777–786. Retrieved from. http://proceedings.mlr.press/v37/foulds15.html
Gal, D., Thijs, B., Glänzel, W., & Sipido, K. R. (2019). Hot topics and trends in cardiovascular research. *European Heart Journal*, *40*(28), 2363–2374. http://dx.doi.org/10.1093/eurheartj/ehz282
Gatti, C. J., Brooks, J. D., & Nurre, S. G. (2015). *A historical analysis of the field of OR/MS using topic models* ArXiv:1510.05154 [Cs, Stat]. Retrieved from. http://arxiv.org/abs/1510.05154
GenomeCanada. (2015). *Genome Canada celebrates 15 years of research and innovation | Genome Canada* Retrieved July 29, 2019, from. https://www.genomecanada.ca/en/news/genome-canada-celebrates-15-years-research-and-innovation
Godin, B. (2003). *The impact of research grants on the productivity and quality of scientific research [Working Paper]*. Ottawa: INRS.
Government of Canada. (2001). *Achieving excellence: Investing in people, knowledge and opportunity—Canada's innovation strategy* Retrieved from. http://publications.gc.ca/collections/Collection/C2-596-2001E.pdf
Government of Canada. (2017). *Investing in Canada's future: Strengthening the foundations of Canadian research* Retrieved from. http://www.sciencereview.ca/eic/site/059.nsf/vwapj/ScienceReview_April2017.pdf/$file/ScienceReview_April2017.pdf
Government of Canada. (2019). *Canada's new superclusters* Retrieved August 28, 2019, from. https://www.ic.gc.ca/eic/site/093.nsf/eng/00008.html

Grajzl, P., & Murrell, P. (2019). Toward understanding 17th century English culture: A structural topic model of Francis Bacon's ideas. *Journal of Comparative Economics*, *47*(1), 111–135. http://dx.doi.org/10.1016/j.jce.2018.10.004
Greenacre, M., Blasius, J., & Blasius, J. (2006). *Multiple correspondence analysis and related methods*. http://dx.doi.org/10.1201/9781420011319
Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(Suppl 1), 5228–5235. http://dx.doi.org/10.1073/pnas.0307752101
Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. http://dx.doi.org/10.1093/pan/mps028
Hale, G. E. (2011). In the pipeline" or "over a barrel"? Assessing Canadian efforts to manage U.S. Canadian energy interdependence. *Canadian - American Public Policy*, *76*, 1–44.
Hartigan, J. A. (1975). *Clustering algorithms*. New York, NY: Wiley.
Herzog, A., John, P., & Mikhaylov, S. J. (2018). *Transfer topic labeling with domain-specific knowledge base: An analysis of UK house of commons speeches 1935-2014. ArXiv:1806.00793 [Cs]* Retrieved from. http://arxiv.org/abs/1806.00793
Hulpus, I., Hayes, C., Karnstedt, M., & Greene, D. (2013). Unsupervised graph-based topic labelling using dbpedia. *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, 465–474. http://dx.doi.org/10.1145/2433396.2433454
Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. SAGE Publications.
Kuhn, K. D. (2018). Using structural topic modeling to identify latent topics and trends in aviation incident reports. *Transportation Research Part C, Emerging Technologies*, *87*, 105–122. http://dx.doi.org/10.1016/j.trc.2017.12.018
Kulczycki, E., Korzeń, M., & Korytkowski, P. (2017). Toward an excellence-based research funding system: Evidence from Poland. *Journal of Informetrics*, *11*(1), 282–298. http://dx.doi.org/10.1016/j.joi.2017.01.001
Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2013). *Handbook of latent semantic analysis*. Psychology Press.
Lastewka, W. (2002). *Canada's innovation strategy: Peer review and the allocation of federal research funds* Retrieved from House of Commons Canada website:. https://www.ourcommons.ca/Content/Committee/371/INST/Reports/RP1032135/indurp10/indurp10-e.pdf
Lau, J. H., Grieser, K., Newman, D., & Baldwin, T. (2011). Automatic labelling of topic models. Proceedings of the 49th annual meeting of the association for computational linguistics. *Human Language Technologies*, *1*, 1536–1545. Retrieved from. http://dl.acm.org/citation.cfm?id=2002472.2002658
Lau, J. H., Newman, D., Karimi, S., & Baldwin, T. (2010). Best topic word selection for topic labelling. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 605–613. Retrieved from. http://dl.acm.org/citation.cfm?id=1944566.1944635
Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*, *23*(2), 254–277. http://dx.doi.org/10.1093/pan/mpu019
Magatti, D., Calegari, S., Ciucci, D., & Stella, F. (2009). Automatic labeling of topics. *2009 Ninth International Conference on Intelligent Systems Design and Applications*, 1227–1232. http://dx.doi.org/10.1109/ISDA.2009.165
Manley, J. (2000). *Report on plans and priorities, 1999-2000 estimates* Retrieved from Natural Sciences and Engineering Research Council (NSERC) website:. http://publications.gc.ca/collections/Collection/BT31-2-2000-III-39E.pdf
Maskeri, G., Sarkar, S., & Heafield, K. (2008). Mining business topics in source code using latent dirichlet allocation. *Proceedings of the 1st India Software Engineering Conference*, 113–120. http://dx.doi.org/10.1145/1342211.1342234
Mehdad, Y., Carenini, G., Ng, R. T., & Joty, S. (2013). Towards topic labeling with phrase entailment and aggregation. Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics. *Human Language Technologies*, 179–189. Retrieved from. https://www.aclweb.org/anthology/N13-1018
Mei, Q., Shen, X., & Zhai, C. (2007). Automatic labeling of multinomial topic models. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 490–499. http://dx.doi.org/10.1145/1281192.1281246
Millar, J. R., Peterson, G. L., & Mendenhall, M. J. (2009). Document clustering and visualization with latent dirichlet allocation and self-organizing maps. *Twenty-Second International FLAIRS Conference. Presented at the Twenty-Second International FLAIRS Conference*,. Retrieved from. https://www.aaai.org/ocs/index.php/FLAIRS/2009/paper/view/62
Mimno, D., & McCallum, A. (2012). *Topic models conditioned on arbitrary features with dirichlet-multinomial regression* ArXiv:1206.3278 [Cs, Stat]. Retrieved from. http://arxiv.org/abs/1206.3278
Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262–272. Retrieved from. http://dl.acm.org/citation.cfm?id=2145432.2145462
Natural Resources Canada. (2019a). *Electricity facts* Retrieved August 28, 2019, from. https://www.nrcan.gc.ca/electricity-facts/20068
Natural Resources Canada. (2019b). *Energy and the economy* Retrieved August 28, 2019, from. https://www.nrcan.gc.ca/energy-and-economy/20062
NSERC. (2001). *Performance report for the period ending March 31, 2001* Retrieved July 29, 2019, from. http://publications.gc.ca/collections/Collection/BT31-4-55-2001E.pdf
NSERC. (2009). *Natural sciences and engineering research council, Departmental performance report for the period ending March 31, 2009* Retrieved from. http://www.nserc-crsng.gc.ca/_doc/reports-rapports/nserc_performance_report_2009_eng.pdf
NSERC. (2015). *How NSERC establishes updated target areas and research topics* Retrieved July 29, 2019, from Natural Sciences and Engineering Research Council of Canada (NSERC) website:. http://www.nserc-crsng.gc.ca/Professors-Professeurs/RPP-PP/SNGNewTargets-SRSNouveauDomaines_eng.asp
NSERC. (2018). *NSERC - departmental plan—2017-18* Retrieved August 23, 2019, from Natural Sciences and Engineering Research Council of Canada (NSERC) website:. http://www.nserc-crsng.gc.ca/NSERC-CRSNG/Reports-Rapports/drr/2017-2018/index_eng.asp
NSERC. (2017). *Collaborative research and training experience (CREATE) program* Retrieved July 29, 2019, from Natural Sciences and Engineering Research Council of Canada (NSERC) website:. http://www.nserc-crsng.gc.ca/Professors-Professeurs/Grants-Subs/CREATE-FONCER_eng.asp
NSERC. (2019). *Dashboard—Natural sciences and engineering research council of canada* Retrieved August 28, 2019, from. http://www.nserc-crsng.gc.ca/db-tb/index-eng.asp?province=0&category=4
NSERC. (2017). *NSERC – Strategic partnership grants* Retrieved July 29, 2019, from Natural Sciences and Engineering Research Council of Canada (NSERC) website:. http://www.nserc-crsng.gc.ca/Professors-Professeurs/RPP-PP/SPG-SPS_eng.asp
NSERC. (2019). *NSERC's awards database* Retrieved from. http://www.nserc-crsng.gc.ca/ase-oro/index_eng.asp
OECD. (2018). *Research and development (R&D)—Gross domestic spending on R&D - OECD Data* Retrieved August 23, 2019, from TheOECD website:. http://data.oecd.org/rd/gross-domestic-spending-on-r-d.htm
Papadimitriou, C. H., Raghavan, P., Tamaki, H., & Vempala, S. (2000). Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, *61*(2), 217–235. http://dx.doi.org/10.1006/jcss.2000.1711
Park, J.-H., & Song, M. (2013). *A study on the research trends in library & information science in Korea using topic modeling.* http://dx.doi.org/10.3743/KOSIM.2013.30.1.007
Paull, R., Wolfe, J., Hébert, P., & Sinkula, M. (2003). Investing in nanotechnology. *Nature Biotechnology*, *21*(10), 1144. http://dx.doi.org/10.1038/nbt1003-1144
Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.
Roberts, M. E., Stewart, B. M., & Tingley, D. (2014). Stm: R package for structural topic models. *Journal of Statistical Software*, https://doi.org/doi:10.18637/jss.v000.i00.
Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., . . . & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082. http://dx.doi.org/10.1111/ajps.12103
Robitaille, J. P., & Gingras, Y. (1999). *The level of funding for university research in Canada and the United States: Comparative study*. Association of Universities and Colleges of Canada.
Rosner, F., Hinneburg, A., Röder, M., Nettling, M., & Both, A. (2014). *Evaluating topic coherence measures* Retrieved from. https://arxiv.org/abs/1403.6397v1

Savoy, J. (2013). Authorship attribution based on a probabilistic topic model. *Information Processing & Management*, *49*(1), 341–354. http://dx.doi.org/10.1016/j.ipm.2012.06.003
Shin, K., Choi, H., & Lee, H. (2015). Topic model analysis of research trend on renewable energy. *Journal of the Korea Academia-Industrial Cooperation Society*, *16*(9), 6411–6418. http://dx.doi.org/10.5762/KAIS.2015.16.9.6411
Statistics Canada. (2011). *Gross domestic expenditures on research and development in Canada (GERD), and the provinces: Analysis* Retrieved August 23, 2019, from. https://www150.statcan.gc.ca/n1/pub/88-221-x/2013001/part-partie1-eng.htm
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://dx.doi.org/10.1002/asi.21435
Sun, L., & Yin, Y. (2017). Discovering themes and trends in transportation research using topic modeling. *Transportation Research Part C, Emerging Technologies*, *77*, 49–66. http://dx.doi.org/10.1016/j.trc.2017.01.013
Taddy, M. A. (2012). On estimation and selection for topic models. *Artificial Intelligence and Statistics*, 1184–1193.
Ubfal, D., & Maffioli, A. (2011). The impact of funding on research collaboration: Evidence from a developing country. *Research Policy*, *40*(9), 1269–1279. http://dx.doi.org/10.1016/j.respol.2011.05.023
van den Besselaar, P., Heyman, U., & Sandström, U. (2017). Perverse effects of output-based research funding? Butler's Australian case revisited. *Journal of Informetrics*, *11*(3), 905–918. http://dx.doi.org/10.1016/j.joi.2017.05.016
Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. *Proceedings of the 26th Annual International Conference on Machine Learning*, 1105–1112. http://dx.doi.org/10.1145/1553374.1553515
Weng, J., Lim, E. P., Jiang, J., & He, Q. (2010). Twitterrank: finding topic-sensitive influential twitterers. *Proceedings of the third ACM international conference on Web search and data mining*, 261–270.
Yan, E. (2014). Topic-based Pagerank: Toward a topic-level scientific evaluation. *Scientometrics*, *100*(2), 407–437. http://dx.doi.org/10.1007/s11192-014-1308-5
Yang, H.-L., Chang, T.-W., & Choi, Y. (2018). Exploring the research trend of smart factory with topic modeling. *Sustainability*, *10*(8), 2779. http://dx.doi.org/10.3390/su10082779
Zeng, A., Shen, Z., Zhou, J., Fan, Y., Di, Z., Wang, Y., . . . & Havlin, S. (2019). Increasing trend of scientists to switch between topics. *Nature Communications*, *10*(1), 1–11. http://dx.doi.org/10.1038/s41467-019-11401-8
Zhang, Y., Zhang, G., Chen, H., Porter, A. L., Zhu, D., & Lu, J. (2016). Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research. *Technological Forecasting and Social Change*, *105*, 179–191. http://dx.doi.org/10.1016/j.techfore.2016.01.015