Scientometrics (2016) 107:477–498
DOI 10.1007/s11192-016-1852-2

iSEER: an intelligent automatic computer system for scientific evaluation of researchers

Ashkan Ebadi[1] · Andrea Schiffauerova[1,2]

Received: 11 July 2015 / Published online: 2 February 2016

Abstract Funding is one of the crucial drivers of scientific activities. The increasing number of researchers and the limited financial resources have caused a tight competition among scientists to secure research funding. On the other side, it is now even harder for funding allocation organizations to select the most proper researchers. Number of publications and citation counts based indicators are the most common methods in the literature for analyzing the performance of researchers. However, the mentioned indicators are highly correlated with the career age and reputation of the researchers, since they accumulate over time. This makes it almost impossible to evaluate the performance of a researcher based on quantity and impact of his/her articles at the time of the publication. This article proposes an **i**ntelligent machine learning framework for **sc**i**e**ntific **e**valuation of **r**esearchers (iSEER). iSEER may help decision makers to better allocate the available funding to the distinguished scientists through providing fair comparative results, regardless of the career age of the researchers. Our results show that iSEER performs well in predicting the performance of the researchers with high accuracy, as well as classifying them based on collaboration patterns, research performance, and efficiency.

Keywords Machine learning · Scientific output · Funding · Research performance · Scientific evaluation

✉ Ashkan Ebadi
a_ebad@encs.concordia.ca; ashkan.ebadi@gmail.com

[1] Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC, Canada

[2] Department of Engineering Systems and Management, Masdar Institute of Science and Technology, Abu Dhabi, United Arab Emirates

Introduction

Research grants are known as one of the crucial drivers of scientific activities that can influence the size and efficiency of the Research and Development (R&D) sector and its productivity (Jacob and Lefgren 2011). They can also affect the performance of researchers through providing them with a better access to the research resources (Lee and Bozeman 2005). Policies on R&D activities have evolved over the past 50 years (Elzinga and Jamison 1995; Sanz Menéndez and Borrás 2000). Several studies specifically analyzed different aspects of the relationship between funding and research performance (Ebadi and Schiffauerova 2013). Funding agencies put a lot of efforts on selecting the best candidates for allocating grants, as well as on evaluating the performance of researchers in regards to the amount of funding that they have been receiving. On the other hand, the growing number of researchers worldwide has made the competition for securing the limited financial resources even harder. For example, the contest for receiving research funding is on the rise in Canada, especially among the academic researchers, mainly due to the changes in federal funding policies, lack of university operating budgets, and increasing research costs (Polster 2007). Researchers' demand for funding cannot be fully satisfied by the finite financial capacity of funding agencies. However, the case could be even worse for young researchers, since their senior counterparts are more known within the scientific community that might help them in securing (more) money for the research (Ebadi and Schiffauerova 2015c).

Peer review is the oldest method that has been being used for evaluating researchers and their grant proposals. Most of the funding agencies use a committee of independent researchers to review the researchers' proposals for funding and to select the most appropriate researcher(s) through a competitive process. However, the peer review process has been widely criticized in the literature due to the potential biases since accuracy of the procedure is highly dependent on the selected experts. For example, preferences of peers can affect the final decision, or they can act as a gatekeeper for new research interests as peers may not come into an integrated conclusion (King 1987). Despite the aforesaid drawbacks, the great advantage of peer review process is that the impact of the proposed research is assessed quite easily and accurately (Allen et al. 2009). For this important reason, it has still remained as one of the most popular techniques in scientific evaluation. One way to overcome the limitations of the expert review is combining it with quantitative performance indicators (Butler 2005; Hicks et al. 2004) in order to achieve an accurate and fair evaluation, since it cannot be reliable enough as a single indicator. For this purpose, citations and publications count based indicators are commonly used as quantitative measures of researchers' performance.

One of the reasons that scientists publish their work in the form of scientific papers is that in this way, they can secure their priority in discoveries (De Bellis 2009). According to the review of literature done by Tan (1986), performance evaluation of individual researchers and research departments are in most cases at least partially based on publication counts measures. Due to the relatively easy access to the required data and simplicity of the calculation, publication count measures are still widely used to analyze the productivity of researchers or research institutes (Van Raan 2005). This includes, but is not limited to, using publication counts to a large extent for measuring the productivity of individual researchers as well as the productivity of the departments (e.g. Porter and Umbach 2001; Dundar and Lewis 1998; Creamer 1998a, b; Bell and Seater 1978).

However, publication counts have some drawbacks, e.g. different nature of work in various scientific disciplines (Wanner et al. 1981), which might affect the accuracy of the analysis.

Apart from the rate of publications, papers impact and visibility should be also taken into the consideration in scientific evaluation. Being first introduced by Gross and Gross in (1927), citation count based indicators are commonly accepted as a proxy for the impact of a scientific publication (Gingras 1996). In general, the mentioned metrics count the number of citations received by an article after the date it is published, and papers with higher number of citations are thus assumed to have higher impact. However, due to the several drawbacks of citation counts, they are not considered by some researchers (e.g. Seglen 1992) as a good measure of the impact of publications. For example, articles of famous researchers are likely to be cited more. In addition, a weak work may receive many citations, not because of its quality, but due to an error in methodology or results discussed by other researchers (Okubo 1997). Nevertheless, citation counts have been widely in use as a significance index of the mean impact of a paper, especially at the aggregate level (Gingras 1996). Some examples are, using citation analyses to evaluate the performance of individual researchers (e.g. Garfield 1970), to evaluate the quality of books (e.g. Nicolaisen 2002), or to analyze the performance of researchers in various scientific fields and academic departments (e.g. Buss 1976).

Publication and citation counts based measures have been widely used for research evaluation in the forms of bibliometrics or statistical analysis (e.g. McAllister and Narin 1983; Peritz 1990; Payne and Siow 2003; Huffman and Evenson 2005; Jacob and Lefgren 2011). The ease of use and suitability of the available databases for applying bibliometric indicators (Luukkonen-Gronow 1987) are some of the reasons for their common use in scientific evaluation. However, such indicators are faced with some limitations in assessing scientific activities, e.g. narrow scope of the study, simplified assumptions and/or indicators. The scenario is almost the same for statistical analyses where in most of the cases a simplified input–output analysis is performed. Recent progress in information technology and the availability of large scale highly accurate integrated digital data have brought new opportunities for scientific evaluation. Powerful computers and complex algorithms have made it possible to come up with new and more accurate solutions.

Scientific collaboration is also one of the important drivers of research progress that supports researchers in generating novel ideas, and influences their scientific activities. Different aspects of collaboration have been studied in a vast number of different disciplines such as computer science, sociology, research policy, and philosophy (Sonnenwald 2007). Several studies assessed the impact of collaboration patterns and network positions on scientific activities and performance of the researchers (e.g. Eslami et al. 2013; Beaudry and Allaoui 2012; Abbasi et al. 2011), and found a positive relation in most of the cases which highlights the influencing role of collaboration in stimulating scientific activities. In addition, through scientific collaboration researchers can get involved in new research projects which might enable them to get access to new financial resources and thus, might result in higher research performance (Ebadi and Schiffauerova 2015d).

Machine learning systems, in particular, have attracted the attention of data analysts in various scientific fields and applications such as stock market, health systems, credit scoring, fraud detection, etc. The ability of automatic learning from data in large scale, instead of manual data manipulation and analysis, has made these potent and modern techniques attractive, not only throughout the computer science field, but also in many other data driven research studies. Hence, machine learning and data mining are expected to be the drivers of the next wave of innovation (Manyika et al. 2011). A very limited number of studies used machine learning techniques in the field of scientific evaluation for predicting the number of citations. Fu and Aliferis (2010) used machine learning techniques to predict the number of citations of biomedical publications. They used support vector machine (SVM) algorithm to learn the input data which contained a number of bibliometric features. In a recent study, Fu et al. (2013) proposed a computer system for identifying the instrumental citations in biomedical publications. They used bibliometric and content based features to train a supervised machine learning model. Their results suggest a high accuracy of the proposed model in classifying the instrumental citations.

In this article, we employ machine learning techniques and propose an integrated framework, named iSEER, for predicting the performance of researchers, as well as their deserved level of funding. We had two main motivations for applying machine learning techniques in scientific evaluation: (1) Since machine learning algorithms are highly data driven, they will often result in more accurate solutions. Moreover, they can be applied in large scope high dimensional data analysis projects and benefit from the variety of the features, as well as the richness of the data, to provide highly accurate tailored solutions, and (2) We believe machine learning and automatic evaluation can help decision makers as a complementary tool that makes the final decisions, with regard to the funding allocation and performance evaluation, less subjective. A shorter version of this paper was presented in the 15th International Conference on Scientometrics and Informetrics (ISSI) in July 2015 (Ebadi and Schiffauerova 2015a). The remainder of the paper proceeds as follows: The next section presents the data and methodology; "Results" section presents the performance evaluation results of iSEER; "Conclusion" section concludes; and limitations and some directions for the future work are presented in the last section.

Data and methodology

Data

The data for this research was collected in five phases. Funding has a determinant role in scientific activities. It is expected that past funding not only affects the current activities of a researcher, but also his/her future level of research money. Hence, as the first phase, we selected Natural Sciences and Engineering Research Council (NSERC) of Canada as the source of funding data. NSERC is the main federal funding organization of the country, covering almost all the Canadian researchers in natural sciences and engineering (Godin 2003). In addition, one of our other motivations for such selection was the availability of NSERC funding data to the public. Moreover, full names of researchers (both first and last names) are listed in NSERC that helped us to perform the entity disambiguation, which will be explained later in this section. Therefore, the NSERC funding data was extracted and stored in a database in the first phase. Several preprocessing and cleaning modules were coded in JAVA and were applied on the collected data to improve the quality of the data. For example, the special characters in the data (e.g. French characters) were automatically detected, and corrected. We also removed students[1] from the funding data, as our purpose was to focus on professional researchers. In addition, for the team grants, the funding amount was equally divided between the principal investigator (PI) and all the co-researchers who were mentioned in the same record. To validate this assumption, we held more than 30 interviews with researchers in our funding database, who were randomly selected through a stratified sampling method, where almost all of them confirmed the

[1] The NSERC database originally contains both scholarships and grants.

assumption of the equal funding division between PI and co-researchers. The resulted funding database contains 379,891 records of researchers, including 102,452 PIs. Since we were interested in the total amount of annual funding for each distinct researcher in the database, we further aggregated the funding database by merging the records for a given researcher in a given year, and adding up the amounts. This step made the set of (researcher, year) unique for each year. The final funding database contains 228,417 records of funded researchers, including 41,024 distinct researchers, within the period of 1996–2010. The funded researchers received the total amount of $18,934,771,899 within the examined period.

Information about the researchers' publications was required for us to be able to assess their performance. In the second phase, Elsevier's Scopus[2] was used to collect researchers' publication data for the period of 1996–2010, including but not limited to the title of the article, co-authors, year of publication and annual citations. Since the data coverage of Scopus was better after 1996, we focused on 1996–2010 time interval. We only collected articles in which NSERC support was acknowledged.[3] For this purpose, a list of keywords (different formats of the way NSERC can be written and spelled) was used as the input and a full text search was automatically performed on the articles. This filtration was a crucial step. The common procedure in the literature is extracting all the articles that were published by a given researcher. Such method suffers from an over-estimation of a given researcher's number of publications since researchers usually have several sources of funding at a time. Our procedure was based on the assumption that NSERC grantees acknowledge the source of funding in their article. NSERC policies and regulations require researchers to mention the source of funding in their publications. To validate this assumption, we held more than 30 interviews with randomly selected funded researchers where almost all of them confirmed that they do acknowledge NSERC in their articles. After performing various automatic data processing and cleaning stages, e.g. parsing affiliations and correcting special characters, the final publication database contains 144,156 distinct articles where 7056 researchers published only one paper. The papers, on average, were cited 2.8 times within the examined time interval.

Having collected the funding and publications databases, the third phase was integrating the mentioned datasets. One of the most important challenges here was matching records in publication database with researchers in NSERC funding database. Since different names and formats were used in Scopus publication data, the entities, i.e. different authors, should have been identified. We faced with two particular problems: (1) To verify whether "Alan Smith", "A. Smith", "A. J. Smith", and "Alan J. Smith" are all pointing to the same person in funding and publication databases or not, and (2) To find out whether "Alan Smith" who is affiliated with the University of Toronto is the same author as the one at McGill University. A JAVA program was coded to perform this crucial task. We had the advantage of the availability of the clean and complete names in NSERC funding dataset, as well as current and past affiliations of the authors, available in Scopus database. Using machine learning methods, the coded semi-automatic JAVA program employed a similarity measure, based on various factors such as names of researchers and their affiliations, to identify and detect entities. We decided to go with a semi-automatic design as the entity

[2] Scopus is a commercial database of scientific articles that has been launched by Elsevier in 2004. It is now one of the main competitors of Thomson Reuter's Web of Science.

[3] We developed a new data extraction methodology which involves a combined use of Google Scholar and Scopus. The main idea is to use the full text search available in Google Scholar and then to search the results in the Scopus database and collect the target articles.

disambiguation task is highly complicated. Thus to minimize the error margins, the program asked user to confirm if the records match, for the cases with the similarity score lower than a pre-defined threshold. The integrated database contains 174,773 records of disambiguated researchers within the period of 1996–2010.

We decided to include two measures for the visibility and impact of publications: (1) Citation based measure which was collected in the second phase, and (2) Rank of journals in which the articles were published. Both measures reflect the impact of publications with a minor difference. Citations based indicators show the impact of the publication on the scientific community and on the subsequent research, whereas journal impact factor or journal rank indicators reflect the respectability of the journal, that is the visibility and the level of contribution perceived by the authors and the reviewers of the paper. Therefore, the fourth phase was dedicated to collecting the journal ranking information for which we selected SCImago Journal Rank (SJR). SCImago was chosen for two main reasons. First, it provides annual data of journal ranks that enabled us to perform a more accurate analysis, since we considered the rank of the journal in the year that an article was published, and not its impact in the current year. Second, SCImago is powered by Scopus that makes it more compatible with our articles database. The collected information was added to the integrated database.

In the final phase of the data gathering procedure, Pajek[4] was used to construct the co-authorship networks of the collected authors for each year of the selected time interval. For this purpose, two-mode co-authorship networks (De Nooy et al. 2011) of authors were first constructed in which both articles and authors are present as the network nodes (Fig. 1a). Next, the constructed two-mode networks were converted to one-mode networks in which two given authors are connected to each other if they have jointly published an article (Fig. 1b). Four network structure variables, i.e. betweenness centrality, degree centrality, clustering coefficient, and eigenvector centrality, were calculated at the individual level of researchers, for each of the authors in the created one-mode networks. The calculated measures were added to the final database. In the final database, we considered only the records for which all the selected measures were available. The size of the final database is 117,942 records. In the next section, the methodology, variables and proposed intelligent framework are introduced and discussed in detail.

Methodology and models

One of the characteristics of iSEER is that it considers various influencing factors of different types and performs the evaluation at the individual level of researchers. As depicted in Fig. 2, the feature space includes variables representing funding, collaboration pattern among researchers, and profile of researchers, as well as their performance. These selected variables are provided to iSEER where the model is trained and outputs are generated. In particular, iSEER covers two types of machine learning models, one for classifying researchers based on their research performance, and the other one for predicting their number of publications as well as their deserved level of funding. In the rest of this section, we will further discuss the mentioned models and their variables.

[4] Social network analysis software, for more information see: http://vlado.fmf.uni-lj.si/pub/networks/pajek/.

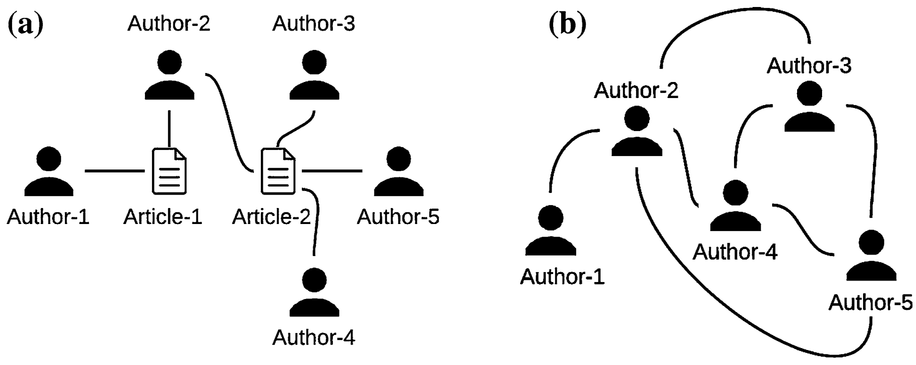

Fig. 1 **a** A sample two-mode co-authorship network. **b** The converted one-mode co-authorship network

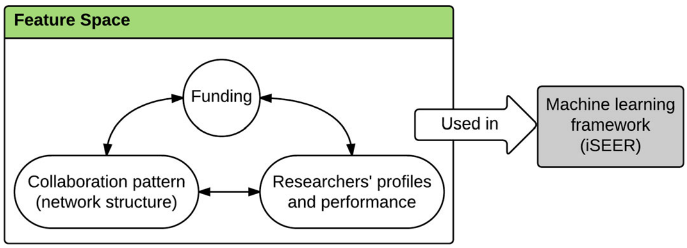

Fig. 2 General schema of the solution

Classification of researchers

Classification is categorizing a new instance (in our case, a new researcher) based on a labeled training dataset. Therefore, we should have a correctly labeled dataset to be able to train the model based on that, and to identify the label (category) of the given data. iSEER performs three types of classification:

- Classifying researchers based on their research performance, i.e. quantity and impact of the papers (Task C1)
- Classifying researchers according to their efficiency (Task C2)
- Classifying researchers based on their rate of collaboration (Task C3)

The only difference in aforementioned tasks is in calculating and assigning the label. To perform Task C1, a label was generated based on both quantity and impact of researchers' publications in a 3-year time window. For this purpose, various indicators and different weights for quantity and impact of the papers were tested. The final research performance indicator, with the most robust results, has three levels (i.e. low, normal, and high performance) in which a relatively higher weight was given to the visibility of the papers, i.e. number of citations and the impact of the journal. The same approach was taken for Tasks C2 and C3. Efficiency of the researchers (Task C2) was evaluated by calculating the cost of article indicator for each of researchers in the database, and by comparing it with the average cost. The final label contains three levels representing low, normal, and high efficiency. For calculating researchers' collaborative behavior index (Task C3), as explained earlier, several combinations were tested where finally the measure was calculated based on the degree centrality (will be defined later in this section) and the average number of co-authors of researchers in a 3-year time window. This label has also three levels reflecting low, normal, and high collaborative behavior of researchers. All the labels were automatically calculated and generated by a JAVA program.

A number of bibliometric features were used as the input to the classification model. High-performing researchers are generally expected to improve (or at least maintain) their performance level. Apart from personal characteristics, one reason is that productive researchers might work on relatively more research projects which will result in higher number of publications. Moreover, they have on average better access to financial resources that can affect their performance and collaboration pattern in future. Therefore, we included past performance measures, in terms of both quantity and impact of publications, in the classification model. Moreover, funding is known as one of the main drivers of scientific activities that can be used by researchers to expand their current activities, to get involved in more projects, to find new partners, to purchase the required equipment, etc. Hence, past funding record of researchers was also added to the model. Different scientific disciplines follow different collaboration patterns as well as various funding allocation procedures. In addition, publication and citation habits might be also different in various scientific fields. For example, citing habits and the rate of citations may vary across different scientific fields such that in some scientific fields authors publish articles more frequently, or the published papers contain more references (MacRoberts and MacRoberts 1996; Phelan 1999). As another example, a lower productivity, in terms of number of publications, is expected from engineers as they are also involved in some other activities, e.g. engineering design (Gingras 1996). Or, in humanities most of the papers are single-authored while in engineering most of the papers have more than one author. In order to stand for such variations, scientific field of researchers was also added to the model. To detect the research domain of the funded researchers, we coded a program implementing Latent Dirichlet Allocation (LDA) technique[5] to extract keywords from the title of the articles, and to categorize articles, and therefore their authors, based on the topics of the articles. We then checked and refined the automatically generated categories. This resulted in 8 different categories, i.e. engineering, mathematics, natural sciences, social sciences, art, health, applied sciences, others. The extracted categories were converted to numerical values, ranging from 0 to 7, to be used in iSEER.

Researchers play different roles in their surrounding and global collaboration networks. These roles can bring various advantages to researchers (e.g. better access to knowledge sources, political factors, and awareness of potential projects) that might enhance or harm their scientific performance, as well as their level of funding. To account for these effects, we included three network structure indicators, i.e. betweenness centrality, clustering coefficient, and degree centrality, in the model. Betweenness centrality is a more global network measure which focuses on the role of intermediary nodes (researchers) in a network and is defined based on the role that a node plays in the existence of paths between any two other nodes as follows (Borgatti 2005):

$$bc_k = \sum_{i \neq k \neq j} \frac{\sigma_{ij}(k)}{\sigma_{ij}} \quad (1)$$

[5] Machine learning technique for topic modeling, first introduced by Blei et al. (2003).

In Eq. (1), σ_{ij} is the total number of shortest paths from node i to j, and $\sigma_{ij}(k)$ is the number of shortest paths from node i to node j that contains node k. Researchers with high betweenness centrality can bridge different communities, control the flow of information and have higher control over the other researchers in the network, in terms of setting project priorities, or knowledge diffusion. Therefore, we expected betweenness centrality to play an important role in scientific activities, thus, it was included in the model.

In graph theory, degree of a given node is calculated as the number of ties that the node has (Diestel 2005). Degree centrality of node i is defined based on the node i degree, where the values are normalized between 0 and 1 as follows:

$$dc_i = \frac{\text{degree of node } i}{\text{highest degree in the network}} \tag{2}$$

In our co-authorship network, researchers with high degree centrality can be more active as they have higher number of direct connections (Wasserman 1994). In addition, in co-authorship networks, degree centrality can be regarded as a proxy for the number of direct collaborators of a researcher. Having more direct collaborators might facilitate the researcher's access to diverse sources of skills and complementary expertise which will make him/her more productive, or may affect his/her level of funding. Therefore, we expect this measure to play a determinant role in scientific activities, hence, it was added to the model.

Clustering coefficient shows the tendency of the nodes to form a cluster together and counts the number of triangles in a given undirected graph to measure the level of clustering. Therefore, it is in fact the likelihood that two neighbors of a node are also connected to each other (Hanneman and Riddle 2011). Theoretically, clustering coefficient is defined based on a local clustering coefficient (*lcc*) for each node within a network. The definition of *lcc* is (Watts and Strogatz 1998):

$$lcc_i = \frac{\text{number of triangles connected to node } i}{\text{number of triples centered on node } i} \tag{3}$$

The denominator in Eq. (3) counts the number of set of two edges that are connected to the node i (triples). The numerator counts the number of three nodes that are all connected to each other. The overall clustering coefficient is calculated by taking average of the local clustering coefficient of all the nodes within the network. Hence,

$$CC = \frac{\sum_{i=1}^{n} LCC_i}{n} \tag{4}$$

In Eq. (4), n denotes the number of vertices in the network. This measure returns a value between 0 and 1 such that it gets closer to 1 as the network interconnectivity increases (higher cliquishness). In co-authorship networks, researchers with high clustering coefficient form tightly connected clusters which might enable them to produce higher quality works through the tight inter-connections in their groups and using the internal referring among the team members. Clustering can also affect the rate of publications. It was shown in the literature that cliquishness might affect the rate of publications negatively hence limiting knowledge creation (e.g. Eslami et al. 2013), which might be due to the exchange of redundant information between closely related communities/clusters (Cowan and Jonard 2003). Hence, we included this measure in the model as well.

After selecting the variables, we needed to take into consideration the time effect. In the literature, 3-year (e.g. Payne and Siow 2003; Beaudry and Allaoui 2012) or 5-year (e.g. Jacob and Lefgren 2011) time windows have been considered for the funding to take effect.

In addition, co-authorship networks, and therefore collaboration patterns, evolve over time. This evolution might reflect the growth/decay of a research subject, community, or even a scientific field (Huang et al. 2008). We tested both 3-year and 5-year time windows for the variables in our model and found better results for the 3-year time window. Hence, to account for the time effect, we considered a 3-year time window for all the selected variables. For example, to assess the performance of a given researcher in 1999 his/her amount of funding was summed up over 1996–1998. The complete list of variables is shown in Table 1.

Figure 3 shows the entire classification process in iSEER for all the above mentioned tasks, i.e. task C1, C2 and C3. As seen, data is first preprocessed and cleaned. For this purpose, several JAVA programs were coded to check the data for redundancy, out of range values, impossible combinations, errors, and missing values, and then the target features (variables) were selected, and data were filtered based on the records that contained all the required data. The resulted data, containing all the potential features, were sent to the data preparation block where at first all the features (except the label) were normalized to a value between 0 and 1. This was a crucial step since the features were of different units and scales. Local Outlier Factor (LOF) algorithm was then applied to detect the outliers. LOF, that was proposed by Breunig et al. (2000), is based on the local density concept in which the local deviation of a given data is measured with respect to its k nearest neighbors. A given data is outlier if it has a substantially different density from its k neighbors. The final step of the data preparation step was optimizing attributes' weights. For this purpose, we used an evolutionary attribute weights optimizer that employed genetic algorithm to calculate the weights of the attributes. The weighting procedure also helped us in detecting the most influential attributes. The resulted data were integrated into a single data repository, named as the target data.

After making the data ready for the analysis, a stratified 10-fold cross validation design was used for the model validation. Cross validation is an analytics tool that is used to design and develop fine-tuned models. It splits the data into two disjoint sets, where one part is used for training and fitting a model (training set), while the other part is employed for estimating the error rate of the model, i.e. test set (Weiss and Kulikowski 1991). We used a nested 10-fold cross validation in which the data were split into 10 disjoint subsets such that the union of the 10 folds results the original data. The method was run 10 times and in each time, one fold was considered as the test data while the rest were regarded as the training data. C4.5 decision tree algorithm (Quinlan 1993) was applied as the model, where its parameters were automatically optimized inside the validation module. We chose C4.5 method as it easily deals with the noise in the data and can handle both categorical (e.g. scientific fields) and continuous variables. In addition, it is an easy to implement

Table 1 List of variables (features) in the classification model of iSEER

	Variables
1	Scientific area of the researcher
2	Total amount of funding received by the researcher in a 3-year time window
3	Total number of publications of the researcher in a 3-year time window
4	Average number of citations received by the researcher's articles in a 3-year time window
5	Average betweenness centrality for the researcher in a 3-year time window
6	Average degree centrality of the researcher in a 3-year time window
7	Average clustering coefficient of the researcher in a 3-year time window

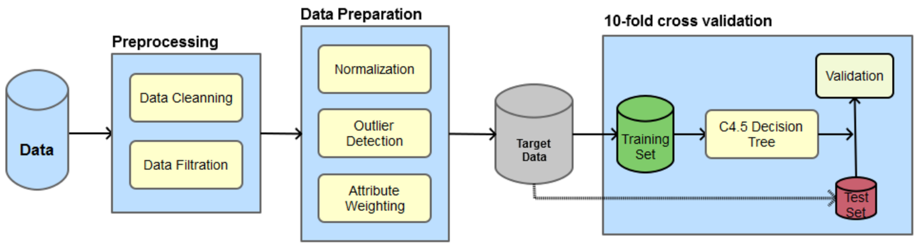

Fig. 3 Classification model, iSEER

method and the results can be easily interpreted, even with limited technical knowledge. In the next section, the prediction models of iSEER are presented.

Prediction of scientific performance and level of funding

We used the same approach as what that was already discussed in the "Classification of researchers" section (classification model) to acquire the target data for the prediction model. Based on the optimized weights, we also added some other variables to the prediction model in comparison with the classification model. Same as the classification model, we used two different proxies for the impact of the papers in the prediction model, i.e. based on citation counts and journal ranks. Age of researchers, and their career level, can influence their performance as well as their funding. It is argued in the literature that older researchers in general can be more productive (Merton 1973; Kyvik and Olsen 2008) due to several reasons, e.g. better access to the funding and expertise sources, more established collaboration network, and better access to modern equipment. Hence, the career age of researchers was included in the model, representing the time difference between the date of their first article in the database and the given year. The average number of co-authors per paper for researchers can be counted as a measure of their average scientific team size, i.e. average number of partners. And, there are several studies that found a positive relation between the team size and scientific output (e.g. Ebadi and Schiffauerova 2015b; Plume and van Wiejen 2014). Thus, this variable was also included in the prediction model, as a common proxy for researchers' collaboration. Apart from the network variables that were already discussed in the "Classification of researchers" section, eigenvector centrality (*ec*) was also added to the prediction model. This centrality measure is based on the idea that the importance of a researcher in the network depends also on the importance of his/her connections. Hence, a researcher with high eigenvector centrality is on average more connected to other researchers, who themselves possess central positions. Bonacich (1972) defined the centrality of an actor based on sum of its adjacent centralities. Being connected to other highly important researchers can bring a strategic and diplomatic power to a researcher which makes the role interesting for our analysis. Such researchers are connected with too many other influential and highly central researchers, and it is hence expected that they shape the collaborations and play an important role in setting priorities in scientific projects, and securing research funding, that might ultimately increase their performance or financial power. The complete list of the final variables for the prediction model is presented in Table 2.

In particular, we defined two prediction tasks in iSEER:

- To predict number of publications of a given researcher (Task P1)
- To predict the deserved amount of funding for a given researcher (Task P2)

To perform the first task (Task P1), we considered the number of publications of researchers as the target variable, while for Task P2, the amount of funding was considered as the target. Figure 4 shows the general scheme of the prediction model. The procedure for preparing the target data is similar to the classification model that was discussed earlier. The difference is in the algorithm, where in the prediction model we used ensemble meta-algorithm to improve the accuracy of the prediction, as we found the prediction task to be more sensitive than the classification task. For this purpose, bootstrap aggregating (bagging) approach was employed. Bagging is an ensemble method that makes random subsets of the data and trains them separately. The final result is then obtained by averaging over the results of the separated models (Breiman 1996). Bagging is a nested module in which we used weighted vote 10-Nearest Neighbor (10-NN) algorithm to train the data and to create the model. In weighted vote 10-NN, the distance of the neighbors to the given data is considered as the weight in the prediction such that neighbors that are closer to the given data get higher weights. The mentioned algorithm was selected for several reasons. First of all, it yielded the highest accuracy among all the candidate algorithms. In addition, it can be easily updated at a very low cost to include new instances. This is a significant advantage for the large databases of publications and authors in real life that should be updated frequently. The ease of implementation and the limited number of parameters that were required to be tuned in this algorithm were some other influencing factors for selecting this model.

Data in the range of 1996–2009 were used to train and build the model. A separate disjoint data for 2010 (prediction set) were used for testing the accuracy of the prediction model. The final result of the prediction model for Task P1 was the predicted number of publications for the researchers in the prediction set. For Task P2, the model calculated a competence factor (between 0 and 1, closer to 1 more competence) that shows the worthiness of a given researcher to receive funding, and used it to predict the amount of funding of a given researcher in 2010. In the next section, results are presented and discussed.

Table 2 List of variables (features) in the prediction model of iSEER

	Attribute
1	Scientific area of the researcher
2	Total amount of funding received by the researcher in a 3-year time window
3	Total number of publications of the researcher in a 3-year time window
4	Average number of citations received by the researcher's articles in a 3-year time window
5	Average rank of the journals in which researcher's' articles were published in a 3-year time window
6	Average betweenness centrality of the researcher in a 3-year time window
7	Average degree centrality of the researcher in a 3-year time window
8	Average clustering coefficient of the researcher in a 3-year time window
9	Average eigenvector centrality of the researcher in a 3-year time window
10	Average scientific team size of the researcher (average number of co-authors per paper)
11	Career age of the researcher

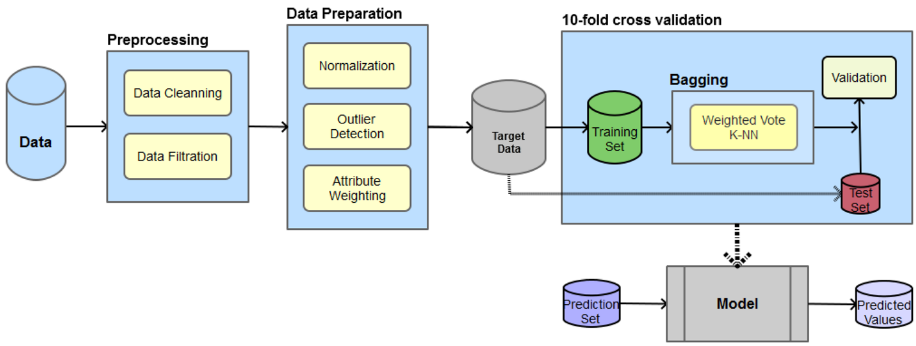

Fig. 4 Prediction model, iSEER

Results

Classification

iSEER was provided with the input data, which was explained in the "Data" section, to evaluate its accuracy in performing the three defined classification tasks, i.e. Tasks C1, C2, and C3. Moreover, we separately tested several machine learning algorithms and compared the accuracy of iSEER with some well-known classifiers. The test results of the top three most accurate algorithms for each task are also listed along with iSEER results. Models were trained and tested on the data from 1996 to 2010. Figure 5 shows the results for the classification tasks. As it can be seen, the accuracy of iSEER is reasonably higher than the other algorithms in performing all the defined classification tasks. Interestingly, apart from 10-NN other classifiers, i.e. Naïve Bayes and decision tree, have considerably lower accuracy than iSEER in performing Task C2. Although Naïve Bayes algorithm is simple and computationally efficient, it is based on strong attribute independence assumptions which might be one of the reasons that this algorithm is not working very well in Task C2 classification. Decision trees are also simple and very easy to understand. However, apart from cost of operation and their complexity, there are some concepts that decision trees cannot learn. Moreover, since our problem is a multi-label classification, the information gain in decision tree can be biased in favor of attributes with higher number of observations (Deng et al. 2011), hence the algorithm might not be able to model the data accurately. This is clearer in the accuracy results of Tasks C1 and C2. The accuracy of iSEER in Task C3 is notably high (98.90 %). Decision Tree comes next in terms of accuracy in Task C3, while 10-NN and Naïve Bayes are coming after it respectively.

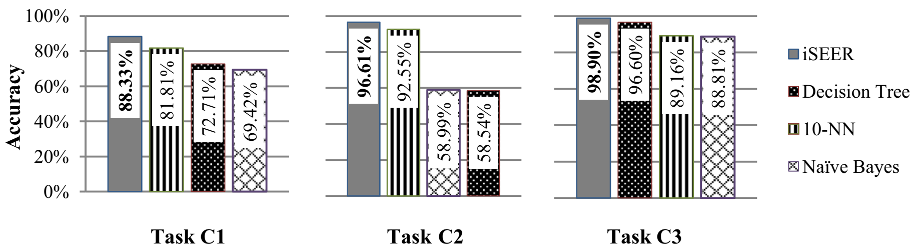

Fig. 5 Accuracy of iSEER versus selected well-known algorithms, classification Tasks C1, C2 and C3

To further evaluate the accuracy of the framework, we compared the confusion matrices of iSEER and the algorithm which has the nearest accuracy to iSEER in different tasks. Confusion matrix was introduced by Kohavi and Provost (1998), and shows the actual and predicted classifications done by a classifier which can be used to evaluate the performance of the classification system. *Precision* and *recall* are two of the measures that are used in the confusion matrix. According to the definition, precision is the proportion of the total number of correct predictions. Recall of a label, in a multi-class problem, is defined as the ratio of correctly predicted cases for that class over the total number of predictions.

As it is seen in Table 3, although iSEER and 10-NN precision and recall are almost comparable for the *predicted high* and *true low* cases in Task C1, iSEER has higher rates of precision and recall in all the sub-classes. For Task C2, precision and recall rates of iSEER is higher than the ones for 10-NN except for the precision of *predicted high* category for which 10-NN is slightly higher. The high accuracy of 10-NN is not very surprising since these classifiers work well when the size of the training data is large. In addition, in our case, we have several features which 10-NN can benefit from to characterize each label based on multiple combinations of the attributes which might increase its accuracy. Analysis of the confusion matrix for Task C3 reveals that iSEER obtained again higher rate of precision than decision tree algorithm except for the *predicted low* category where the difference is almost negligible (99.55 vs. 100 %). For the recall rates, iSEER also performs better except for the *true high* category where the difference is small (96.90 vs. 98.70 %). In general and considering the performance in all the defined tasks, it can be said that iSEER is a more accurate classifier for the subject problem. Moreover, its accuracy in classifying low and high performing researchers is of great importance for considering it as a complementary tool for decision makers. In the next section, we check the performance of iSEER in researchers' evaluation procedure.

Prediction

In this section, we present the test results of iSEER in predicting performance of researchers (Task P1), as well as their deserved amount of funding (Task P2). We trained the model with the data from 1996 to 2009 and used the disjoint set of data for 2010 for

Table 3 Confusion matrix of iSEER versus the best performing algorithm in Tasks C1, C2 and C3

	Precision			Recall		
	Predicted low (%)	Predicted normal (%)	Predicted high (%)	True low (%)	True normal (%)	True high (%)
Task C1						
iSEER	94.28	78.51	84.59	94.36	79.69	82.53
10-NN	87.74	67.61	83.79	92.88	67.78	68.53
Task C2						
iSEER	98.03	94.78	96.58	97.75	95.89	94.85
10-NN	92.99	90.11	97.14	96.53	89.72	87.06
Task C3						
iSEER	99.55	98.32	97.17	99.70	98.16	96.90
Decision tree	100	92.77	91.48	96.15	96.66	98.70

Fig. 6 Accuracy of iSEER versus selected well-known algorithms, prediction Tasks P1 and P2

predicting the target variables and testing the accuracy. We also compared the accuracy of the model with several well-known machine learning algorithms, where in this section, the test results of iSEER, along with two other algorithms that showed the highest accuracy in predicting the target variable in each task, are presented. Figure 6 shows iSEER prediction errors in both tasks compared with other algorithms.

We considered three error measures for comparing the performance of the mentioned algorithms. Root mean squared error is one of the main measures for comparing the accuracy of the prediction models and is defined as the square root of the average of the squares of errors. According to Fig. 6, iSEER is predicting the number of publications of the researchers (Task P1) with 1.451 average deviation between the predicted value and the real number of publications. Normalized absolute error is the absolute error, i.e. difference between the predicted value and the real value, divided by the error made if the average would have been predicted. The root relative squared error takes the average of the actual values as a simple predictor to calculate the total squared error. The result is then normalized by dividing it by the total squared error of the simple predictor and square root is taken to transform it to the same dimension as the predicted value. As seen in Fig. 6, iSEER performs the best, according to all the three measures, where the polynomial fit is the worst in performing Task P1. Results for Task P2 are slightly different where linear regression and 10-NN algorithms were the two closest algorithms to iSEER, in terms of the prediction errors. According to Fig. 6, root mean squared error of iSEER is the lowest in Task P2 where the other two algorithms perform the same. Although linear regression normalized absolute error is a bit lower than iSEER, its root relative squared error surpasses iSEER. Hence, according to the results, it can be claimed that the overall performance of iSEER is slightly better than the other algorithms. A sample of the prediction results is presented in Appendix A.

Conclusion

In this paper, we proposed iSEER system that uses bibliometric indicators as well as network structure features to classify researchers based on their collaboration patterns, research performance, and efficiency. It can also predict the number of publications of researchers along with their deserved amount of research funding. According to our results, it is feasible to employ machine learning algorithms for classification of the researchers based on various criteria. Moreover, it was shown that iSEER can predict the performance and deserved funding level of researchers, with relatively high accuracy. In addition, the unique procedure that was presented in this research highlighted the most important features in classifying researchers, as well as the ones in predicting their performance.

As discussed before, iSEER is able to automatically predict the amount of funding as well as a normalized competence factor (between 0 and 1) for a given researcher. Since in real life the amount of money is finite, the normalized competence factor can act as a complementary coefficient, helping decision makers to set the final amount of funding for a given researcher. The predicted funding level can be also used for comparing with the final allocated amount, or for making intra-researchers comparisons. The predicted number of publications explicitly shows the expected performance of a researcher in coming year(s), which might be also helpful. Regarding the classification tasks, various applications can be considered such as, using them in selecting the best candidate for a vacant research or academic position.

Although few researchers recently worked on citation prediction using machine learning algorithms (e.g. Fu and Aliferis 2010; Fu et al. 2013), to our knowledge iSEER is the first system that focuses on research performance and funding prediction, as well as classifying researchers, using various features of different types, e.g. bibliometric and collaboration network structure indicators. The intensive preprocessing steps along with feature selection procedure, helped iSEER to achieve high predictive power and accuracy rate. The result of attribute weighting module also shed light on influential attributes in predicting or categorizing the target researchers. Moreover, several features of similar nature were employed in the model to reinforce its accuracy. For example, we used average number of citations and average rank of the journals to represent the visibility of the papers. Another example is the use of degree centrality and average number of authors per paper, to represent the scientific team size of researchers. These attributes of similar nature surely empowered the accuracy of the model by providing it with additional dimensions.

To conclude, our results show that it is feasible to design and use classification and prediction tools to evaluate different aspects of scientific activities of researchers. It is obvious that peer reviewing cannot be completely replaced by such tools. iSEER can help decision makers in setting both long-run and short-term strategies in regard to the funding allocation and/or analyzing researchers' performance and scientific collaboration patterns among the researchers through providing them with more accurate quantitative analysis. In addition, since our framework is flexible, and high dimensional data and a large dataset was used for learning the model, the results are not based on limited criteria or data. Therefore, it can also help decision makers to establish a fairer funding allocation or scientific evaluation system. Lastly, we believe the field of scientific evaluation can benefit from the advancement in computer science in at least three ways as presented in this research: (1) Sophisticated and well-tailored data gathering procedure(s) can definitely provide the analysts with more accurate data in a very large scale, which can help them to better analyze the inter-relations; (2) Complex computer algorithms, in form of intelligent automatic systems, can be used to perform more accurate quantitative analysis; (3) Data mining and machine learning can also serve as a tool for selecting the important factors (variables) in a study, no matter what evaluation method is used afterwards.

Limitations and future work

We were exposed to some limitations in this paper. First, we selected Scopus for gathering information about the NSERC funded researchers' publications. Scopus and other similar databases are English biased thus non-English articles are underrepresented (Okubo 1997).

Since Scopus data coverage was better after 1996, we chose the time interval of 1996–2010 for our analysis. Although Scopus is confirmed in the literature to have a good coverage of articles, as a future work it would be recommended to focus on other similar databases to compare and confirm the results.

Furthermore, we were exposed to some limitations in measuring scientific collaboration among the researchers, as we were unable to capture other links that might exist among the researchers, e.g. informal relationships. These types of connections are never recorded and thus cannot be quantified, but there are certainly some knowledge exchange occurring during such associations that could affect the network performance. In addition, there are also some drawbacks in using co-authorship as an indicator of scientific collaboration since collaboration does not necessarily result in a joint article (Tijssen 2004). An example could be the case when two scientists cooperate together on a research project and then decide to publish their results separately (Katz and Martin 1997). Hence, future work can address this issue by taking other types of collaboration networks into the consideration. For assessing the impact of the papers based on citations count, we did not account for self-citations, negative citations, or special inter-citation patterns among a number of researchers. This can be addressed in future works. In addition, we used citation based indicators along with SJR journal rank for representing and assessing the impact of research. Of course, other respective variables can be added to the model to measure the impact more accurately. That is, iSEER can be easily expanded to include more influencing factors, or to be tailored for other scientific domains/projects.

Appendix: Sample of prediction results

Variables are listed in Table 4 and samples of the predictions for both prediction tasks are presented in Tables 5 and 6. The real value of the target variable is highlighted in light grey where the respective predicted value is highlighted in dark grey.

Table 4 Description of the variables

Variable	Description
Discip	Scientific area of the researcher
sumFund3	Total amount of funding received by the researcher in a 3-year time window
noArt3	Total number of publications of the researcher in a 3-year time window
avgCit3	Average number of citations received by researcher's articles in a 3-year time window
avgIf3	Average rank of the journals in which researcher's articles were published in a 3-year time window
btwn3	Average betweenness centrality of the researcher in a 3-year time window
deg3	Average degree centrality of the researcher in a 3-year time window
clust3	Average clustering coefficient of the researcher in a 3-year time window
eigen3	Average eigenvector centrality of the researcher in a 3-year time window
teamSize	Average number of authors per paper for the researcher
careerAge	Career age of the researcher

Table 5 Sample of iSEER prediction results, Task P1

Predicted no articles	noArt	sum Fund3	avg If3	avg Cit3	teamSize	btwn3	clust3	deg3	eigen3	careerAge	discip	noArt3
0.361	0	0.041	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.737	2	0
1.102	0	0.013	0.279	0.028	0.000	0.000	1.000	0.005	0.000	0.632	3	1
3.865	7	0.044	0.054	0.005	0.001	0.059	0.125	0.027	0.000	0.737	1	13
1.103	0	0.010	0.068	0.083	0.000	0.000	1.000	0.007	0.000	0.737	3	1
1.206	1	0.072	0.132	0.020	0.002	0.016	0.409	0.020	0.000	0.526	0	6
6.703	4	0.167	0.246	0.080	0.002	0.055	0.158	0.039	0.000	0.737	1	26
1.030	4	0.032	0.115	0.017	0.001	0.018	0.455	0.018	0.000	0.737	0	6
4.120	3	0.061	0.136	0.041	0.002	0.185	0.109	0.134	0.000	0.737	1	15
0.000	0	0.012	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.263	0	0
5.047	3	0.137	0.141	0.041	0.001	0.133	0.163	0.050	0.000	0.684	0	15

Table 6 Sample of iSEER prediction results, Task P2

Predicted Fund	sum Fund	sum Fund3	avg If3	avg Cit3	team Size	btwn3	clust3	deg3	eigen3	careerAge	discip	noArt3
$414,936	$53,515	0.205	0.189	0.092	0.002	0.008	0.222	0.009	0.000	0.579	1	0.096
$70,832	$69,786	0.023	0.141	0.010	0.002	0.000	0.600	0.005	0.000	0.474	1	0.019
$60,750	$51,880	0.011	0.132	0.019	0.002	0.000	0.444	0.008	0.000	0.737	2	0.019
$183,301	$239,331	0.072	0.150	0.042	0.001	0.016	0.409	0.011	0.000	0.526	0	0.058
$78,938	$49,918	0.023	0.178	0.019	0.000	0.001	0.500	0.004	0.000	0.684	1	0.019
$158,689	$159,600	0.073	0.140	0.010	0.001	0.007	0.400	0.005	0.000	0.526	1	0.019
$131,313	$114,421	0.042	0.096	0.070	0.002	0.048	0.257	0.014	0.000	0.737	0	0.077
$117,806	$88,280	0.043	0.101	0.029	0.001	0.001	0.333	0.004	0.000	0.737	0	0.019
$85,018	$58,800	0.022	0.080	0.019	0.001	0.000	0.000	0.001	0.000	0.368	0	0.010
$74,211	$106,750	0.017	0.051	0.074	0.001	0.000	1.000	0.004	0.000	0.105	0	0.019

References

Abbasi, A., Altmann, J., & Hossain, L. (2011). Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics, 5*(4), 594–607.

Allen, L., Jones, C., Dolby, K., Lynn, D., & Walport, M. (2009). Looking for landmarks: The role of expert review and bibliometric analysis in evaluating scientific publication outputs. *PLoS One, 4*(6), e5910.

Beaudry, C., & Allaoui, S. (2012). Impact of public and private research funding on scientific production: The case of nanotechnology. *Research Policy, 41*(9), 1589–1606.

Bell, J. G., & Seater, J. J. (1978). Publishing performance: Departmental and individual. *Economic Inquiry, 16*(4), 599–615.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993–1022.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology, 2*(1), 113–120.

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks, 27*(1), 55–71.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.

Breunig, M. M., Kriegel, H., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *ACM Sigmod Record, 29*(2), 93–104.

Buss, A. R. (1976). Evaluation of Canadian psychology departments based upon citation and publication counts. *Canadian Psychological Review/Psychologie Canadienne, 17*(2), 143.

Butler, L. (2005). *What happens when funding is linked to publication counts? Handbook of quantitative science and technology research* (pp. 389–405). NewYork: Springer.

Cowan, R., & Jonard, N. (2003). The dynamics of collective invention. *Journal of Economic Behavior & Organization, 52*(4), 513–532.

Creamer, E. G. (1998). *Assessing faculty publication productivity: Issues of equity. ASHE-ERIC higher education report, volume 26, number 2*. ERIC.

Creamer, E. G. (1998). *Assessing faculty publication productivity: Issues of equity. ASHE-ERIC higher education report, volume 26, number 2*. ERIC.

De Bellis, N. (2009). *Bibliometrics and citation analysis: From the science citation index to cybermetrics*. Lanham: Scarecrow Press.

De Nooy, W., Mrvar, A., & Batagelj, V. (2011). *Exploratory social network analysis with Pajek* (27th ed.). Cambridge: Cambridge University Press.

Deng, H., Runger, G., & Tuv, E. (2011). Bias of importance measures for multi-valued attributes and solutions. *Artificial neural networks and machine Learning–ICANN 2011* (pp. 293–300) Springer.

Diestel, R. (2005). Graph theory, vol. 173 of. *Graduate Texts in Mathematics*.

Dundar, H., & Lewis, D. R. (1998). Determinants of research productivity in higher education. *Research in Higher Education, 39*(6), 607–631.

Ebadi, A., & Schiffauerova, A. (2013). Impact of funding on scientific output and collaboration: A survey of literature. *Journal of Information & Knowledge Management, 12*(04), 1350037.

Ebadi, A., & Schiffauerova, A. (2015a). A computer system for automatic evaluation of researchers' performance, *15th International Society of Scientometrics and Informetrics Conference (ISSI 2015)*, 425–435.

Ebadi, A., & Schiffauerova, A. (2015b). On the relation between the small world structure and scientific activities. *PLoS One, 10*(3), e0121129.

Ebadi, A., & Schiffauerova, A. (2015c). How to receive more funding for your research? Get connected to the right people! *PLoS One, 10*(7), e0133061.

Ebadi, A., & Schiffauerova, A. (2015d). How to become an important player in scientific collaboration networks? *Journal of Informetrics, 9*(4), 809–825.

Elzinga, A., & Jamison, A. (1995). Changing policy agendas in science and technology. In Sheila Jasanoff (Ed.), *Handbook of science and technology studies*. London: Sage.

Eslami, H., Ebadi, A., & Schiffauerova, A. (2013). Effect of collaboration network structure on knowledge creation and technological performance: The case of biotechnology in Canada. *Scientometrics, 97*(1), 99–119.

Fu, L. D., & Aliferis, C. F. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics, 85*(1), 257–270.

Fu, L. D., Aphinyanaphongs, Y., & Aliferis, C. F. (2013). Computer models for identifying instrumental citations in the biomedical literature. *Scientometrics, 97*(3), 871–882.

Garfield, E. (1970). Citation indexing for studying science. *Essays of an Information Scientist, 1*, 133–138.

Gingras, Y. (1996). *Bibliometric analysis of funded research. A feasibility study*. Ottawa: Report to the Program Evaluation Committee of NSERC.

Godin, B. (2003). The impact of research grants on the productivity and quality of scientific research. No. 2003. *INRS working paper.*

Gross, P. L. K., & Gross, E. M. (1927). College libraries and chemical education. *Science, 66*(1713), 385–389.

Hanneman, R. A., & Riddle, M. (2011). Concepts and measures for basic network analysis. In J. Scott, & P. J. Carringon (Eds.), *The SAGE handbook of social network analysis* (pp. 340–367). Thousand Oaks, CA: Sage.

Hicks, D., Tomizawa, H., Saitoh, Y., & Kobayashi, S. (2004). Bibliometric techniques in the evaluation of federally funded research in the United States. *Research Evaluation, 13*(2), 76–86.

Huang, J., Zhuang, Z., Li, J., & Giles, C. L. (2008). Collaboration over time: Characterizing and modeling network evolution. *Proceedings of the 2008 international conference on web search and data mining*, pp. 107–116.

Huffman, W. E., & Evenson, R. E. (2005). New econometric evidence on agricultural total factor productivity determinants: Impact of funding composition. *Iowa State University*, *Department of Economics*, *Working paper*, *3029.*

Jacob, B. A., & Lefgren, L. (2011). The impact of research grant funding on scientific productivity. *Journal of Public Economics, 95*(9), 1168–1177.

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy, 26*(1), 1–18.

King, J. (1987). A review of bibliometric and other science indicators and their role in research evaluation. *Journal of Information Science, 13*(5), 261–276.

Kohavi, R., & Provost, F. (1998). Confusion matrix. *Machine Learning, 30*(2–3), 271–274.

Kyvik, S., & Olsen, T. B. (2008). Does the aging of tenured academic staff affect the research performance of universities? *Scientometrics, 76*(3), 439–455.

Lee, S., & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science, 35*(5), 673–702.

Luukkonen-Gronow, T. (1987). Scientific research evaluation: A review of methods and various contexts of their application. *R&D Management, 17*(3), 207–221.

MacRoberts, M. H., & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics, 36*(3), 435–444.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

McAllister, P. R., & Narin, F. (1983). Characterization of the research papers of US medical schools. *Journal of the American Society for Information Science, 34*(2), 123–131.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago: University of Chicago press.

Nicolaisen, J. (2002). The J-shaped distribution of citedness. *Journal of Documentation, 58*(4), 383–395.

Okubo, Y. (1997). *Bibliometric indicators and analysis of research systems: Methods and examples* (Vol. 1997/1). Paris: OECD Publishing.

Payne, A. A., & Siow, A. (2003). Does federal research funding increase university research output? *Advances in Economic Analysis & Policy*, *3*(1), 1–22.

Peritz, B. C. (1990). The citation impact of funded and unfunded research in economics. *Scientometrics, 19*(3–4), 199–206.

Phelan, T. (1999). A compendium of issues for citation analysis. *Scientometrics, 45*(1), 117–136.

Plume, A., & van Wiejen, D. (2014). Publish or perish? The rise of the fractional author. *Trends Journal of Sciences Research,* 38.

Polster, C. (2007). The nature and implications of the growing importance of research grants to canadian universities and academics. *Higher Education, 53*(5), 599–622.

Porter, S. R., & Umbach, P. D. (2001). Analyzing faculty workload data using multilevel modeling. *Research in Higher Education, 42*(2), 171–196.

Quinlan, J. R. (1993). *C4. 5:* Programs for machine learning, *Morgan Kaufmann.*

Sanz Menéndez, L., & Borrás, S. (2000). Explaining changes and continuity in EU technology policy: The politics of ideas.

Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science, 43*(9), 628–638.

Sonnenwald, D. H. (2007). Scientific collaboration. *Annual Review of Information Science and Technology, 41*(1), 643–681.

Tan, D. L. (1986). The assessment of quality in higher education: A critical review of the literature and research. *Research in Higher Education, 24*(3), 223–265.

Tijssen, R. J. (2004). Is the commercialisation of scientific research affecting the production of public knowledge?: Global trends in the output of corporate research articles. *Research Policy, 33*(5), 709–733.

Van Raan, A. F. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics, 62*(1), 133–143.

Wanner, R. A., Lewis, L. S., & Gregorio, D. I. (1981). Research productivity in academia: A comparative study of the sciences, social sciences and humanities. *Sociology of Education, 54*, 238–253.

Wasserman, S. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440–442.

Weiss, S., & Kulikowski, C. (1991). *Computer systems that learn*. California: Morgan Kaufmann Publishers.