Research at the Frontier of Knowledge: Comparing text similarity indicators to citations for measuring scientific excellence

Roman Fudickar and Hanna Hottenrott

December 2020

Abstract
Evaluating scientific excellence is a fundamental challenge for public science administrations. It currently primarily relies on peer-review and publication data analysis. This study proposes to add text-based indicators to the evaluation procedure, in order to get a more comprehensive picture of a scientists potential to do excellent science. We compare text-based similarity between publications of individual scientists in different scientific fields (biology, chemistry, economics and engineering) and text-documents of validated knowledge frontiers to citation-based indicators. We propose two knowledge frontiers for science evaluations (academic prizes and ERC funding) and show that text similarity approaches can be a valuable complement to standard bibliometric indicators. Moreover, survey data is used to study their relationship with alternative individual-specific measures of research quality, such as academic rank, institutional rank, and research budget. We find that overall text-based, citation-based and survey-based indicators provide a coherent picture. However, for young researchers for whom citations windows are short, text-based indicators may provide additional insights when evaluating research excellence. Moreover, the correlation between similarity scores and citation measures decreases with scientists' age indicating their use also for established researchers.

Keywords: scientific excellence, research evaluation, natural language processing
JEL Codes: I20, O30, O38

Contact details:
hanna.hottenrott@tum.de (corresponding author), TUM School of Management, Technische Universität München, Arcisstraße 21, 80333 München, Germany
r.fudickar@tum.de, TUM School of Management, Technische Universität München, Arcisstraße 21, 80333 München, Germany

Acknowledgements: We thank participants of the 23rd international conference on Science and Technology Indicators, participants of the Max-Planck Institute for Innovation and Competition Brownbag Seminar, and Henry Sauermann for helpful comments. This research was funded by the German Research Foundation (DFG), grant number HO 5390/1-1.

1. Introduction

Identifying scientific excellence is crucial for public science administrations that aim to allocate scarce research funding to the most promising projects and researchers. This study presents an analysis of established and new indicators of scientific excellence. In particular, it addresses the research question whether text-based similarity between publications of individual scientists in different scientific fields (i.e. biology, chemistry, economics and engineering), and documents of validated knowledge frontiers can serve to evaluate scientific excellence. Specifically, we compare citation counts and text-based similarity indicators, to study their relationship with alternative subject-specific measures of research quality, such as academic rank, institutional rank, and research budget. The comparison aims to validate whether text-based, citation-based and research related indicators provide a coherent picture, or if they point in perpendicular directions.

For the construction of indicators, we calculate document-document similarity scores between a sample of 1884 scientists and two knowledge frontier definitions. The first frontier definition is based on 575 recent science prize awardees and their scientific publications between 2011 and 2016. The second knowledge frontier is based on the project descriptions of 3114 prestigious research grants (European Research Council grants) awarded during the same time period. Both knowledge frontiers involve a highly competitive peer-review process, are based on recent achievements in advancing human knowledge, and thus appear suitable as a reference point of excellence in science. The underlying text data was obtained from the publication records of each sample author and each prize awardee. We merged each authors' titles, keywords and abstracts into one document per author (henceforth sample documents, frontier documents). For ERC projects, we downloaded the project information from the EU CORDIS database and combined title and project objective into one document per project. After this, we used common text mining techniques like filtering, tokenization, and term weighting to standardize the vocabulary for the comparison. After pre-processing, we used co-word analysis to obtain similarity scores between each sample document and each frontier document in the respective field using four binary and four metric similarity measures. This co-word analysis resulted in 16 average similarity scores per sample author.

The results show that such average similarity scores correlate highly with citations and other individual-level indicators of research quality. Higher similarity scores, namely

authors with a higher proximity to the knowledge frontier, have higher research budgets, more senior academic ranks, and work for higher ranked institutions. Furthermore, we find that the correlation between similarity scores and citation measures decreases with scientists' age. This study studies the feasibility of text-based similarity scores for science evaluations that aim to identify excellent scientists. Furthermore, we propose the utility of two knowledge frontiers for science evaluations (academic prizes and ERC funding). Although text similarity may not reflect a similar scientific alignment (or quality) in a variety of cases, we argue that given the "right" reference points, pre-processing and parameters - text similarity approaches can be valuable to complement peer review and standard bibliometric indicators, especially for evaluating younger researchers who lack an long publication record for citation analyses.

2. Background and purpose of the study

In addition to peer review and citation counts, esteem indicators inform quality assessments. Esteem indicators[2] are non-bibliometric indicators of research quality which are rather based on the standing of an individual or of pieces of research within the academic community. Previous research suggests that such indicators can provide an external "reference point" or a knowledge frontier, to which other scientists can be compared (e.g. Frey and Neckermann 2008, 2010; Zuckerman 1992). However, according to Frey and Neckermann (2008), there is "almost no serious empirical evidence on the effects of awards on (research) performance, mainly because the properties and effects of awards have rarely been studied by economists or by other social scientists." (p. 5).

Similar to academic awards, prestigious research grants can reflect excellence. Funding is essential for a scientist's work, and it contributes substantially to successful research outcomes (Stephan 1996; Hottenrott and Lawson 2017). Financial support from prestigious funding institutions signals the ability to undergo competitive peer-review processes through excellent research ideas and vindication. In this study, we propose that reputable research grants can also be assessed as a signal for scientific excellence. A fundamental difference of research funding to academic prizes, which are awarded in retrospect, is that research funding is awarded to scientists who commit to do great

[2] Esteem indicators include honours, awards and prizes; election to academies and academic professional associations; service to conferences or journals; visiting fellowships, or prestigious research finding.

research in the future. Thus, prestigious funding may also serve as a benchmark to identify scientific excellence which has not been addressed in previous studies.

Content-based analysis of scientific communication is a promising research avenue that allows evaluators to overcome certain shortcomings of peer evaluation and citations numbers. Content analysis of publications appears to be more scalable, cheaper, faster, and less prone to evaluation bias than peer evaluation. However it can only mechanically identify some of the complex, dynamic and often subtle patterns of research excellence. Content analysis of scientific publications allows to find patterns which are invisible for citation analyses. One of such patterns is text similarity between publications of individual scientists and documents of validated knowledge frontiers. In this study, we therefore explore the feasibility and plausibility of content-based indicators which are based on frontier knowledge for science evaluation. Specifically, we address the research question of whether text-based similarity between publications of individual scientists and publications of award or ERC grant winners can be an indicator for research excellence. To answer this question, we investigate the correlation of content-based similarity scores with citation counts and a set of research quality indicators (i.e. research budgets, academic ranks and institution ranks) that usually associate with research performance.

3. Data

The first knowledge frontier definition is based on international academic prize awardees. We identify relevant academic prizes using Zheng and Lius' (2015) list of "important international academic awards"[3] (see Appendix I for details). From this list, we take all available prizes in four focal disciplines to identify recent price recipients. In particular, we include 10 prizes in economics, business and finance; 34 prizes in life sciences (biology and biosciences and medicine); 11 in chemistry; and 54 in engineering to our study.[4] We then looked up the recipients' names for the five past award periods (usually annual or biannual recognition) and their respective Scopus identification numbers

[3] These awards are selected on three criteria: a) They honour individuals' contribution to the advancement of knowledge (i.e. research awards); b) that are not restricted "on nationality, and generally regardless of race, gender, age, religion, ethnicity, sexual orientation, disability, language, or political affiliation"; and are c) "generally granted by international organizations, national governments, renowned foundations, academic associations, national academies and learned societies".

[4] We exclude Nobel Prize winners, since they are typically awarded for life-time academic achievements rather than for recent academic achievements.

(Scopus ID). After manual cleaning and disambiguation of names and affiliations, we downloaded all publication records which listed the researcher as an author from the Scopus database for the time period 2011-2016. If more than one Scopus ID for a given author was found, we merged their records into one document. We retain only peer-reviewed English language articles for the period 2011-2016. From the list of prize winners, we further remove the duplicate entries of those scientists who won more than one award in a discipline during this period. We also exclude those authors that did not have peer-reviewed articles in the focal time period and publications without either abstract or keywords. This selection resulted in 575 prize awardees of which 45% are active in engineering, 37% in biology and medicine, and 9% in each chemistry and economics or business. We then combine all available titles, keywords and abstracts of frontier authors into single text documents (*frontier documents*$_{pri}$).

The second frontier definition is based on grants awarded by the European Research Council (ERC). The ERC is the most prestigious European funding organization with the aim to support long-term funding of curiosity-driven research at the frontiers of knowledge. ERC grants are designed to support high-risk basic research and pioneering research without topical restriction. The selection of grantees is conducted by peer review panels composed of renowned scientists, with "scientific excellence" being the principal selection criterion. We consider 3664 projects that were granted between 2011 and 2016 and which were tagged by at least one subject area.[5] This resulted in project descriptions for 1897 starting grants, 313 consolidator grants, 1430 advanced grants, and 24 synergy grants. We downloaded their project information from the EU Horizon 2020 framework website (CORDIS) and merged the project title and description into single text documents (*frontier documents*$_{erc}$).

For information on a focal group of researchers, we use data of individual researchers collected by the "International Science Affiliations" project conducted at the Technical University of Munich in 2016. The sampled authors were randomly chosen from journals stratified by their eigenfactor score in four scientific disciplines: biology (27%), chemistry (31%), economics and business (20%), and engineering (23%); see Appendix II for descriptive statistics and a description of the survey. We complemented the survey

[5] Some projects have multiple field tags what causes the reported sum in the collection statistics table (A.6) to be larger.

data by downloading each respondent's publication records from the Scopus database until 2016. The publications were restricted to English language articles in peer-reviewed academic journals. Further, we added the number of co-authors per publication to control for team size effects (Persson et al. 2004). In the following, we refer to these scientists as *sample authors*. In Appendix III, we provide an overview of the twelve collections used in this study.

3.1 Method and procedures

We use co-word analysis to calculate the scientific proximity between sets of sample and frontier documents. Scientific proximity is a spatially visualized representation of how fields, subjects, publications and authors are related, based on ideational or cognitive proximity (in contrast to physical proximity, Small 1999, see Appendix IV for an illustration).[6] Co-word analysis is a text mining technique that extracts words from documents, standardizes the vocabulary and builds a matrix of word co-occurrences between documents (see Appendix V for details on the method and the key parameters). We further map the frontier fields from the academic prizes and the ERC projects to the four sample disciplines, i.e. biology, chemistry, economics and business, and engineering; to only compare authors within their discipline (see Appendix VI for details). Next, we utilize four binary and four metric similarity measures that describe the statistical congruence between author document vectors (see Appendix VII for details).

Procedure

With the pre-processed document vectors and four specified algorithm parameters, we create a term-document matrix for each sample collection in a specific field and the respective frontier document collection (Table 1). Based on these matrices, we calculate average similarity scores between each sample document *i* and each frontier document *j*, for both knowledge frontiers *f*, using eight similarity measures *m* in each scientific domain *d*.

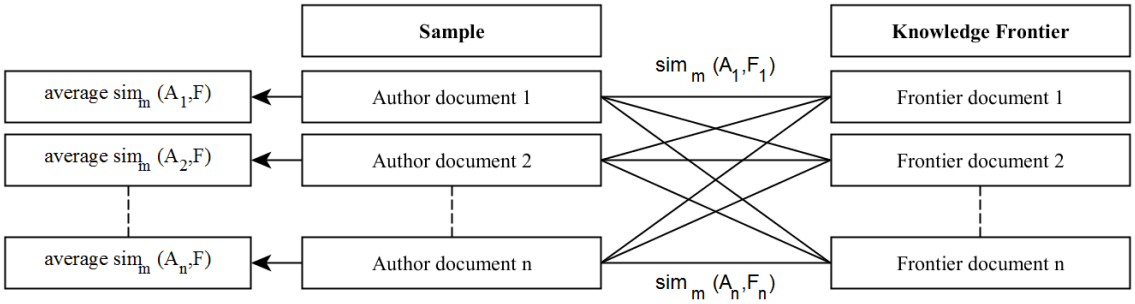
[6] Despite scientific proximity, it exists a considerable body of literature that is concerned with technological proximity based on patent information (e.g. Bar and Leiponen 2012) or hybrid studies of scientific and technological proximity (Magermann et al. 2010).

Table 1: Document collection overview

	Academic prize collections	**ERC collections**
Focal author collections	$sample_{bio}$ *vs.* $prize_{bio}$	$sample_{bio}$ *vs.* $funding_{bio}$
	$sample_{che}$ *vs.* $prize_{che}$	$sample_{che}$ *vs.* $funding_{che}$
	$sample_{eco}$ *vs.* $prize_{eco}$	$sample_{eco}$ *vs.* $funding_{eco}$
	$sample_{eng}$ *vs.* $prize_{eng}$	$sample_{eng}$ *vs.* $funding_{eng}$

Figure 1 illustrates this procedure. From this calculation, we obtain 16 average similarity scores for each of the sample authors (eight measures and two frontiers). We normalize these average similarity scores by setting the highest resulting similarity score to one, the lowest score to zero, and all other scores relative to them (min-max normalization).

Figure 1: Similarity calculation procedure



Average similarity scores

We provide an overview of the resulting similarity scores and a brief characterization of them in Appendix VIII.

Empirical model

We use the obtained average similarity indicators in four basic OLS regression models. These models are identical with respect to the independent and control variables and only vary in the dependent variable. We test whether a) *avg. similarity* $score_{pri}$ b) *avg. smilarity* $score_{erc}$, c) $ln(citations_{total})$ *or* $ln(citations_{per\ article})$ are explained by the same characteristics typically found in excellent scientists. As independent variables, we add categorical variables for each quartile of the research budget (*1st – 4th quartile*), for each academic rank (*junior, post-doc, assistant professor and full professor*), and for each institution rank (*tier1, tier 2, tier 3, not ranked*) to the regression models:

$$research\ quality_i = \beta_0 + \beta_1 research\ budget_i + \beta_2 academic\ rank_i + \beta_3 institution\ rank + \sum_{n=4}^{k} \beta_n\ controls_i + u_i \qquad (1)$$

A set of control variables are included that have been shown to affect publication outcomes, such as age, gender, country, and field (Toole and Czarnitzki, 2010; Mairesse and Pezzoni, 2015). Moreover, we add the number of co-authors per paper as a control to the regression models since documents with many co-authors obtain more citations than documents with fewer co-authors (Persson et al. 2004).

3.2. Findings

Correlation analysis

We compare the average similarity scores with scientists' citation counts, academic rank, institution rank, and annual funding budget. The basic idea of this comparison is to see whether the scores actually correlate with what has previously been related to research quality or excellence. Figure 1 provides several scatter plots which display the relationship between similarity scores and citations per article as logged variables. Most plots show a positive relationship between similarity scores and citations per article and only the simple matching coefficient has a negative correlation.[7] For the academic prize frontier, the explained variance R^2 ranges between 8% (correlation) and 15% (Jaccard, Dice, Cosine, eJaccard and eDice). The explained variance of the score using the funding frontier ranges between 8% (Jaccard and Dice) and 21% (Russel). We repeat these scatter plots with the log of total citations in Figure 3 and find an identical positive relationship between scores and citations, however with a much higher correlation between the scores and total citations, rather than citations per article.

Younger scientists, such as doctoral and postdoctoral students, have a structural disadvantage when their research quality is gauged by citation indicators. This is because citations largely depend on scientific visibility, which junior scientists typically lack. To test this idea for the obtained similarity scores, we analyse the interaction effect of age and the similarity score with respect to citations. From Tables 2 and 3, we find that the interaction effect between score and age is negative and significant for all metric similarity scores. For binary similarity measures, the Jaccard and Dice index are insignificant while the simple matching coefficient and the Russel index are strongly

[7] The negative correlation is potentially the result of the simple matching formula in Table A.10. The formula incorporates d (mutual absence of terms) and normalizes by n (number of unique terms in the whole term-document matrix). Using these auxiliary variables, the similarity scores may not reflect the real congruence between two focal documents, but rather the structure of the TDM, e.g. how many other documents are included in the matrix.

significant (not shown here). This suggests that the correlation between similarity scores and citations is stronger for young scientists and decreases with age, and that text similarity to frontier knowledge can be a valuable substitute when citation counts are less meaningful. The increasing dissimilarity with author's age / seniority may be connected to the growing experience throughout researchers' careers. Young scholars might begin their careers by understanding and replicating established authors and studies, and thus have a high similarity to existing frontier knowledge. When scholars gain more research experience, they also improve in their ability to judge the novelty of ideas and to position themselves in the market for ideas (and thus have a low similarity). Similarly, it is easier for established researchers to attract research funding. This enables them to address more novel (unprecedented) research ideas, which in turn are less similar to existing frontier research ideas.

Table 2: The moderating effect of age (prize frontier)

	ln(citations)			
	cosine	**ejaccard**	**edice**	**correlation**
similarity score	7.078*** (.594)	7.164*** (.598)	7.048*** (.594)	5.193*** (.683)
age	.022*** (.005)	.022*** (.005)	.022*** (.005)	.035*** (.006)
similarity score ## age	-.026** (.012)	-.027** (.012)	-.026** (.012)	-.026* (.014)
_cons	.824*** (.254)	.863*** (.250)	.828*** (.255)	.866*** (.303)
observations	1884	1884	1884	1884
R^2	.49	.49	.49	.29

Notes: *** (**,*) indicate a significance level of 1% (5%, 10%).

Table 3: The moderating effect of age (funding frontier)

	ln(citations)			
	cosine	**ejaccard**	**edice**	**correlation**
similarity measure	7.299*** (.621)	7.250*** (.616)	7.230*** (.616)	6.807*** (.651)
age	.045*** (.007)	.046*** (.007)	.046*** (.007)	.056*** (.007)
similarity measure ## age	-.049*** (.013)	-.049*** (.013)	-.049*** (.013)	-.056*** (.014)
_cons	-.237 (.317)	-.265 (.316)	-.292 (.319)	-.363 (.336)
observations	1884	1884	1884	1884
R^2	.42	.42	.42	.32

Notes: *** (**,*) indicate a significance level of 1% (5%, 10%).

Figure 2: Correlation between similarity scores and citations per article

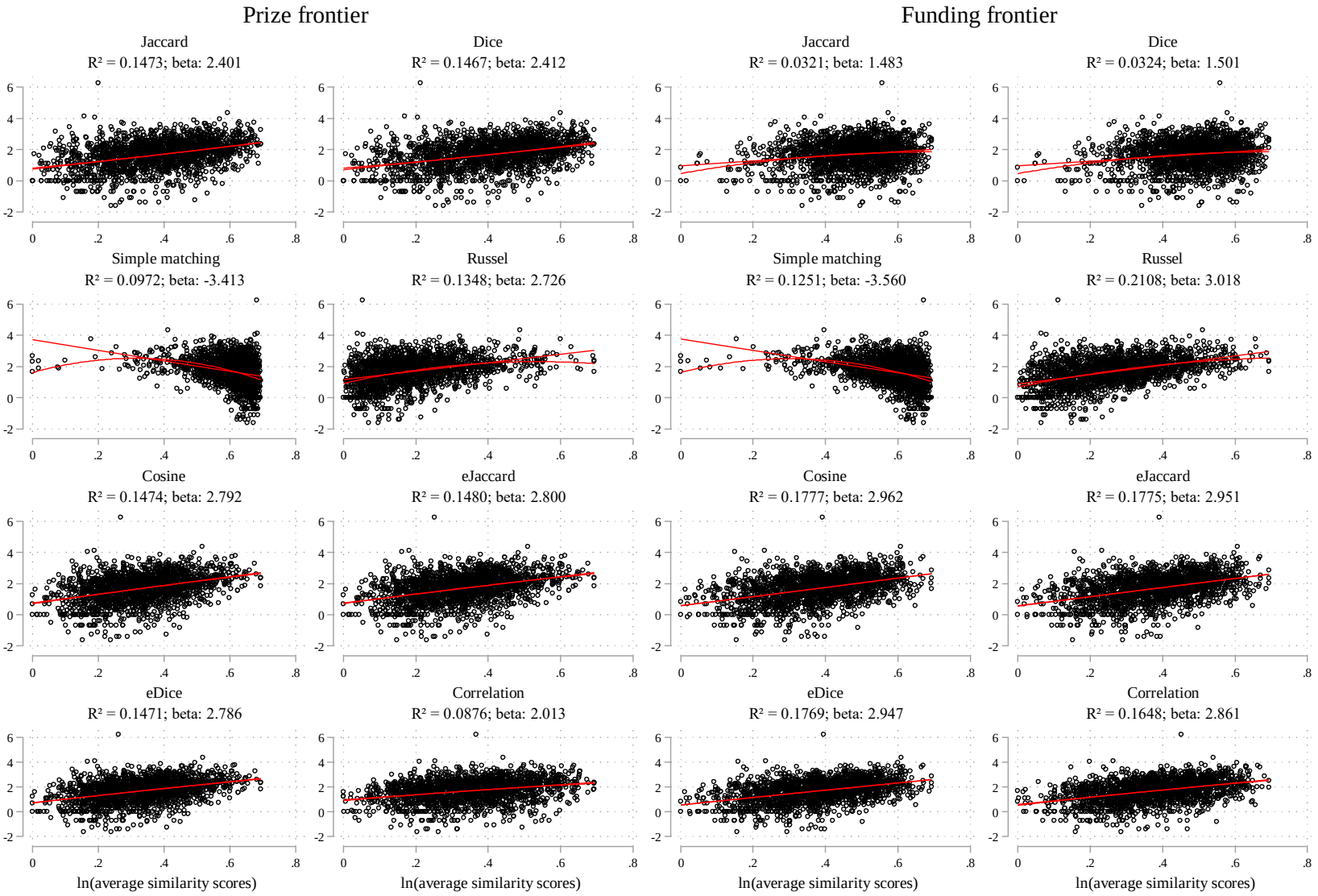
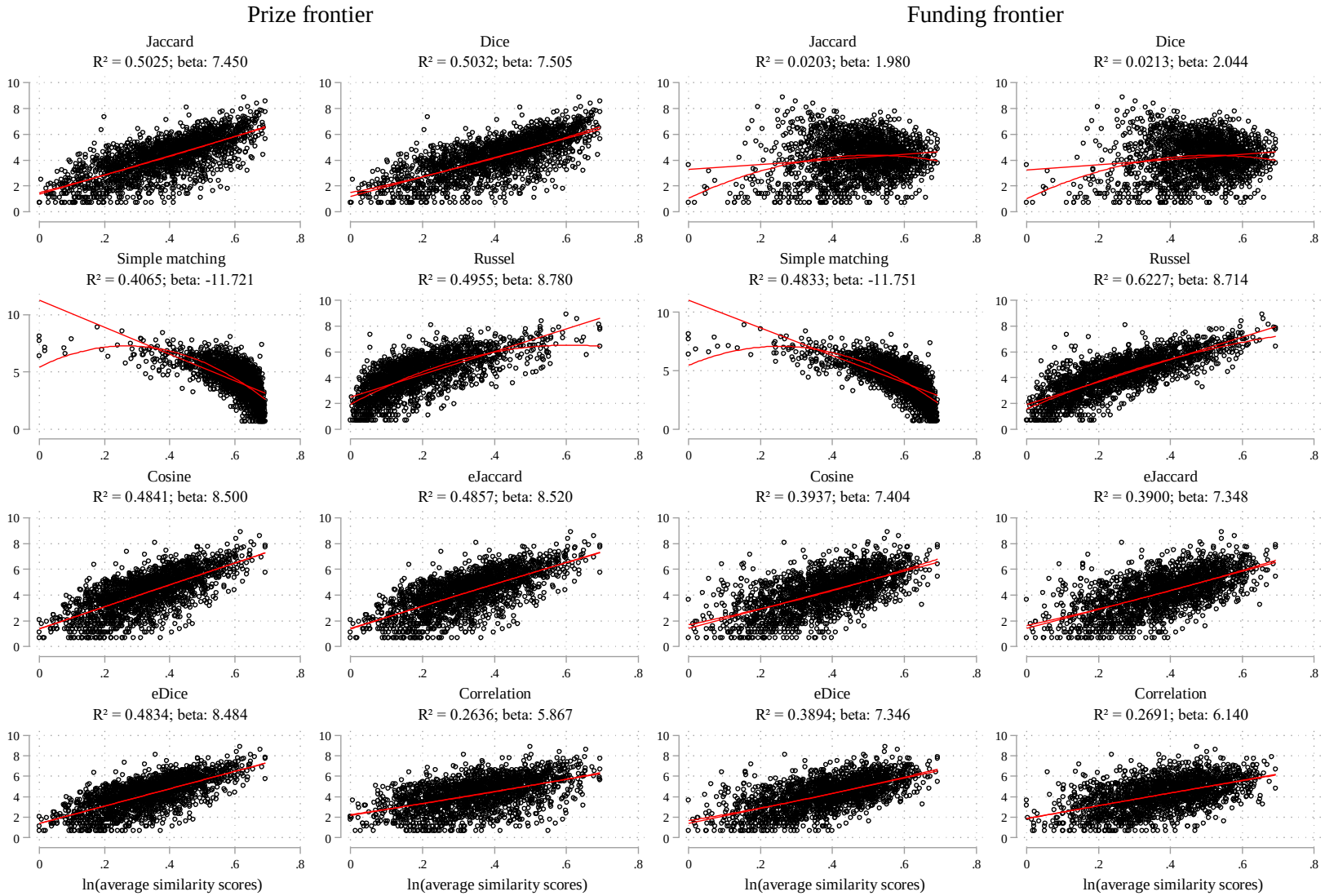


Figure 3: Correlation between similarity scores and citations



In Table 4, we compare citation models to similarity score models for the prize frontier. Scientists with a relatively high research budget (in the 4th quartile) have on average more total citations and more citations per article, than scientists with a very low research budget (1st quantile). We also observe a strong relationship between academic ranks and the considered outcome variables. Full professors and medium ranked scientists like post-docs and assistant professors have on average more citations and more citations per article than their junior colleagues without PhD. A coherent picture is also visible for the institution ranks. Scientists from high and medium ranked institutions (Tier 1 and Tier 2) exhibit higher citation indicators and similarity scores compared to scientists of unranked institutions. Tier 1 scientists have more total citations and more citations per article than scientists from unranked institutions. Overall, we find for the prize frontier that the similarity scores correlate positively with institutional ranks, research budget and institution rank.

In Table 5, we provide the same empirical setting as in Table 4, however with similarity scores obtained for the funding frontier. The first two models are equally specified to those in Table 4 and are only repeated for comparison. For similarity scores based on the funding frontier, we find again that the similarity scores are also positively related to research budgets, academic rank, and institution rank. However the magnitude of the quality indicators varies slightly.

With respect to the control variables, the number of published documents has a positive correlation with citation and similarity indicators. We further see that additional co-authors per publication increase the citation and similarity indicators (as was shown by Persson et al. 2004). Also age has a positive correlation with citations per article and a weak correlation with similarity scores. However, the quadratic age coefficient is negative and significant, indicating that the similarity scores diminish for older scientists. While we do not find correlations between gender and the citation indicators, we find that women have on average lower similarity to both knowledge frontiers.

Overall, we find that three common research quality indicators (research budget, academic rank, and institution rank) show a positive correlation with both sets of citation and similarity indicators. With respect to differences of frontier definitions, we find a remarkable resemblance between the coefficients of the prize and funding frontier. Their coefficients deviate only slightly, tend to be lower for the funding frontier, but do not contradict each other in terms of

algebraic sign. Further, we find that the explanatory power (R^2) of the models using the prize frontier are higher than those of the funding frontier.

Previous studies typically preferred metric over binary measures. While metric measures incorporate the term frequency to allow for term weighting schemes (e.g. TF/IDF), binary measures only utilize the presence or absence of terms thus giving equal weight for each term. Our analysis shows that most included similarity measures are consistent in terms of algebraic sign and distribution. We observe that Jaccard and Dice index (binary), and Cosine, eJaccard, eDice, and Pearson correlation (metric) follow a similar distribution and only slightly deviate from each other in the regression analysis. Only the simple matching and Russel coefficient appear problematic since their formulas use the auxiliary variables d (mutual absence of terms) and n (total number of terms in all documents of the TDM), c.f. Appendix VII. This might cause these measures to not reflect the similarity of two focal documents, but rather quantifying the underlying TDM structure. Therefore, it seems plausible for future studies to explore the consistency of multiple similarity measures.

Table 4: OLS regression (prize frontier)

	ln(citations)	ln(citations per article)	jaccard	dice	simple matching	russel	cosine	ejaccard	edice	correlation
research budget					1st quartile = reference category					
2nd quartile	.042 (.055)	.106 (.068)	.014 (.010)	.015 (.010)	-.001 (.005)	.004 (.006)	.009 (.008)	.009 (.008)	.009 (.008)	.013 (.011)
3rd quartile	-.003 (.057)	.048 (.071)	.028*** (.011)	.028** (.011)	-.003 (.005)	.016** (.006)	.028*** (.009)	.029*** (.009)	.029*** (.009)	.047*** (.012)
4th quartile	.137* (.072)	.346*** (.089)	.070*** (.013)	.072*** (.013)	-.013** (.006)	.027*** (.007)	.046*** (.011)	.045*** (.011)	.046*** (.011)	.052*** (.014)
academic rank					junior = reference category					
postdoctoral position	.279** (.124)	.827*** (.141)	.131*** (.018)	.137*** (.019)	-.034*** (.007)	.053*** (.010)	.098*** (.014)	.097*** (.014)	.099*** (.014)	.109*** (.016)
assistant professor	.317** (.129)	.972*** (.148)	.159*** (.020)	.166*** (.020)	-.038*** (.008)	.062*** (.010)	.118*** (.015)	.116*** (.015)	.119*** (.015)	.135*** (.018)
full professor	.411*** (.134)	1.175*** (.161)	.196*** (.022)	.202*** (.023)	-.055*** (.009)	.090*** (.012)	.150*** (.017)	.148*** (.017)	.151*** (.017)	.165*** (.020)
institution rank					not ranked = reference category					
tier 3	-.010 (.053)	.072 (.067)	.013 (.010)	.013 (.010)	-.003 (.004)	.002 (.006)	.003 (.008)	.003 (.008)	.003 (.008)	-.007 (.010)
tier 2	.094** (.046)	.195*** (.057)	.032*** (.009)	.033*** (.009)	-.009** (.004)	.012** (.005)	.020*** (.007)	.020*** (.007)	.020*** (.007)	.018* (.010)
tier 1	.205*** (.048)	.294*** (.060)	.036*** (.009)	.038*** (.009)	-.008* (.004)	.014** (.006)	.033*** (.007)	.032*** (.007)	.033*** (.007)	.043*** (.010)
publications	.007*** (.001)	.035*** (.003)	.005*** (.000)	.005*** (.000)	-.005*** (.000)	.006*** (.000)	.005*** (.000)	.005*** (.000)	.005*** (.000)	.004*** (.000)
co-authors per article	.059*** (.009)	.090*** (.012)	.014*** (.002)	.014*** (.002)	-.007*** (.001)	.009*** (.001)	.011*** (.002)	.011*** (.002)	.011*** (.002)	.011*** (.002)
age	.010 (.014)	.037** (.017)	.008*** (.003)	.009*** (.003)	-.004*** (.001)	.004*** (.001)	.006*** (.002)	.005*** (.002)	.006*** (.002)	.003 (.003)
age²	-.000 (.000)	-.000** (.000)	-.000*** (.000)	-.000*** (.000)	.000*** (.000)	-.000*** (.000)	-.000*** (.000)	-.000*** (.000)	-.000*** (.000)	-.000** (.000)
female	.052 (.052)	-.055 (.061)	-.019** (.009)	-.019** (.010)	.000 (.004)	-.008 (.005)	-.016** (.007)	-.016** (.007)	-.016** (.007)	-.025** (.010)
country					Japan = reference category					
Germany	.536*** (.062)	.705*** (.080)	.145*** (.012)	.147*** (.012)	-.049*** (.006)	.080*** (.007)	.118*** (.010)	.116*** (.009)	.118*** (.010)	.133*** (.013)
United Kingdom	.634*** (.062)	.749*** (.077)	.126*** (.012)	.127*** (.012)	-.043*** (.005)	.070*** (.007)	.102*** (.009)	.100*** (.009)	.102*** (.009)	.110*** (.013)
field					engineering = reference category					
biology	.393*** (.054)	.406*** (.067)	.004 (.010)	-.009 (.010)	-.005 (.005)	-.010* (.006)	-.047*** (.008)	-.044*** (.008)	-.045*** (.008)	-.115*** (.010)
chemistry	.429*** (.050)	.587*** (.065)	-.054*** (.009)	-.059*** (.010)	.078*** (.004)	-.069*** (.005)	-.090*** (.007)	-.085*** (.007)	-.090*** (.007)	-.121*** (.010)
economics	-.270*** (.066)	-.522*** (.084)	.086*** (.014)	.073*** (.014)	-.039*** (.007)	.065*** (.008)	.009 (.011)	.014 (.011)	.011 (.011)	-0.000364
_cons	.132 (.335)	.454 (.398)	-.169*** (.058)	-.154** (.060)	1.125*** (.024)	-.145*** (.031)	-.046 (.047)	-.053 (.046)	-.047 (.047)	.080 (.061)
observations	1884	1884	1884	1884	1884	1884	1884	1884	1884	1884
R^2	.31	.63	.56	.55	.76	.72	.58	.58	0.58	0.37

Notes: *** (**,*) indicate a significance level of 1% (5%, 10%).

Table 5: OLS regression (funding frontier)

	ln(citations)	ln(citations per article)	jaccard	dice	simple matching	russel	cosine	ejaccard	edice	correlation
research budget					1st quartile = reference category					
2nd quartile	.042 (.055)	.106 (.068)	.017 (.011)	.017 (.011)	-.002 (.005)	.004 (.008)	.009 (.010)	.010 (.010)	.010 (.010)	.006 (.011)
3rd quartile	-.003 (.057)	.048 (.071)	.002 (.011)	.002 (.011)	-.008 (.006)	.021** (.008)	.024** (.011)	.025** (.011)	.025** (.011)	.025** (.012)
4th quartile	.137* (.072)	.346*** (.089)	.051*** (.015)	.051*** (.015)	-.023*** (.006)	.047*** (.010)	.060*** (.013)	.061*** (.014)	.061*** (.014)	.059*** (.015)
academic rank					junior = reference category					
postdoctoral position	.279** (.124)	.827*** (.141)	.068*** (.022)	.069*** (.022)	-.045*** (.008)	.078*** (.014)	.093*** (.018)	.095*** (.018)	.095*** (.018)	.077*** (.018)
assistant professor	.317** (.129)	.972*** (.148)	.079*** (.023)	.079*** (.023)	-.052*** (.009)	.093*** (.015)	.108*** (.019)	.109*** (.019)	.110*** (.019)	.092*** (.020)
full professor	.411*** (.134)	1.175*** (.161)	.079*** (.025)	.080*** (.025)	-.072*** (.011)	.122*** (.017)	.137*** (.021)	.138*** (.021)	.139*** (.021)	.114*** (.022)
institution rank					not ranked = reference category					
tier 3	-.010 (.053)	.072 (.067)	.010 (.010)	.010 (.010)	-.004 (.005)	.005 (.007)	.005 (.010)	.005 (.010)	.005 (.010)	.000 (.011)
tier 2	.094** (.046)	.195*** (.057)	.029*** (.009)	.029*** (.009)	-.012*** (.005)	.023*** (.007)	.032*** (.009)	.032*** (.009)	.032*** (.009)	.029*** (.010)
tier 1	.205*** (.048)	.294*** (.060)	.032*** (.010)	.032*** (.010)	-.011** (.005)	.022*** (.007)	.038*** (.009)	.038*** (.009)	.038*** (.010)	.040*** (.010)
publications	.007*** (.001)	.035*** (.003)	-.001*** (.000)	-.001*** (.000)	-.005*** (.000)	.006*** (.000)	.004*** (.000)	.003*** (.000)	.003*** (.000)	.002*** (.000)
co-authors per article	.059*** (.009)	.090*** (.012)	.003** (.001)	.003** (.001)	-.008*** (.001)	.011*** (.002)	.010*** (.002)	.010*** (.002)	.010*** (.002)	.007*** (.002)
age	.010 (.014)	.037** (.017)	.003 (.003)	.003 (.003)	-.005*** (.001)	.007*** (.002)	.006** (.002)	.006** (.002)	.006** (.002)	.005* (.003)
age²	-.000 (.000)	-.000** (.000)	-.000 (.000)	-.000 (.000)	.000*** (.000)	-.000*** (.000)	-.000*** (.000)	-.000*** (.000)	-.000*** (.000)	-.000** (.000)
female	.052 (.052)	-.055 (.061)	-.020* (.010)	-.020* (.010)	.003 (.005)	-.011* (.007)	-.015* (.009)	-.016* (.009)	-0.000144	-.015 (.010)
country					Japan = reference category					
Germany	.536*** (.062)	.705*** (.080)	.119*** (.012)	.119*** (.012)	-.062*** (.006)	.122*** (.009)	.162*** (.012)	.163*** (.012)	.163*** (.012)	.170*** (.013)
United Kingdom	.634*** (.062)	.749*** (.077)	.151*** (.012)	.151*** (.012)	-.053*** (.006)	.119*** (.009)	.179*** (.011)	.180*** (.012)	.180*** (.012)	.202*** (.012)
field					engineering = reference category					
biology	.393*** (.054)	.406*** (.067)	-.088*** (.010)	-.088*** (.010)	-.008 (.005)	.003 (.007)	-.049*** (.009)	-.060*** (.009)	-.061*** (.009)	-.074*** (.010)
chemistry	.429*** (.050)	.587*** (.065)	-.105*** (.010)	-.107*** (.010)	.074*** (.005)	-.054*** (.007)	-.117*** (.009)	-.117*** (.009)	-.119*** (.009)	-.068*** (.010)
economics	-.270*** (.066)	-.522*** (.084)	-.087*** (.013)	-.087*** (.013)	-.016** (.007)	-.020** (.010)	-.067*** (.012)	-.080*** (.012)	-.081*** (.012)	-.108*** (.013)
_cons	.132 (.335)	.454 (.398)	.483*** (.065)	.486*** (.065)	1.136*** (.027)	-.175*** (.042)	.048 (.057)	.059 (.058)	.063 (.058)	.141** (.062)
observations	1884	1884	1884	1884	1884	1884	1884	1884	1884	1884
R^2	.31	.63	.22	.22	.76	.67	.42	.41	0.41	0.3

Notes: *** (**,*) indicate a significance level of 1% (5%, 10%).

4. Conclusion and Discussion

The prime goal of the analysis was to explore the technical feasibility and plausibility of content-based indicators for identifying scientific excellence of individual scientists. The research question which we addressed is here was how text-based similarity between publications of individual scientists and documents of validated knowledge frontiers compare to citation-based and survey-based indicators.

The results confirm that document-document similarity between individual scientists' publications and knowledge frontier documents appears to capture scientific excellence. We find that four common research quality indicators (i.e. citations, research budget, academic rank and institution rank) show a positive correlation with the derived text similarity indicators. We interpret these findings as some initial evidence for the idea that content-based analyses based on knowledge frontiers can be valuable for science evaluations when citation measures may be less meaningful. This is potentially the case for younger scholars since their citation numbers had less time to accumulate. We suggest that policy makers and administrators may consider such indicators for research funding allocation and science evaluations. Our study shows that content-based indicators are a valuable source of information which can complement peer review and standard bibliometric indicators. Different types of indicators reflect particular dimensions of research quality. A diverse set of indicators may provide evaluators with more valid and more useful assessment tools to estimate scientific excellence. Text similarity to excellent projects and persons may provide one of such indicators, but also dissimilarity could inform evaluators about novel and unprecedented combinations of terms.

A limitation of the proposed method is its insufficiency to fully deal with lexical ambiguity and variability, for example synonymy, antonymy, homonymy, polysemy, acronyms, negations, alternations, abbreviations, etc. (Hotho et al. 2005). Future research might consider other methods that account for lexical ambiguity (e.g. topic models or part-of speech-tagging of technical terms), for content-based analyses in science evaluations.

References

Bar, T., & Leiponen, A. (2012). A measure of technological distance. *Economics Letters*, 116(3), 457-459.

Deza, M. M., & Deza, E. (2009). Encyclopedia of distances. In *Encyclopedia of Distances*, 1-583, Springer Berlin Heidelberg.

Frakes, W. B., & Baeza-Yates, R. (1992). *Information retrieval: Data structures & algorithms* (Vol. 331). Englewood Cliffs, New Jersey: prentice Hall.

Frey, B. S., & Neckermann, S. (2008). Academics appreciate awards-a new aspect of incentives in research. *CESifo Working Paper Series No. 2531*

Frey, B. S., & Neckermann, S. (2010). Awards as signals. *CESifo Working Paper Series No. 3229*

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. In *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, 20(1), 19-62.

Hottenrott, H., & Lawson, C. (2017). Fishing for complementarities: Research grants and research productivity. *International Journal of Industrial Organization*, 51, 1-38.

Lenz, H. J. (2008). Proximities in statistics: Similarity and distance. *Preferences and Similarities*, 161-177.

Magerman, T., Van Looy, B., & Song, X. (2010). Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics*, 82(2), 289-306.

Mairesse, J., & Pezzoni, M. (2015). Does gender affect scientific productivity?. *Revue économique*, 66(1), 65-113.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Persson, O. (1994). The intellectual base and research fronts of JASSIS 1986–1990. *JASSIS*, 45(1), 31–38.

Persson, O., Glänzel, W., & Danell, R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics*, 60(3), 421-432.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130-137.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the Association for Information Science and Technology*, 50(9), 799.

Stephan, P. E. (1996). The economics of science. *Journal of Economic literature*, 34(3), 1199-1235.

Toole, A. A., & Czarnitzki, D. (2010). Commercializing Science: Is there a university “brain drain” from academic entrepreneurship?. *Management Science*, 56(9), 1599-1614.

Zheng, J., &Liu, N. (2015). Mapping of important international academic awards. *Scientometrics*, 104(3), 763-791.

Zuckerman, H. (1992). The proliferation of prizes: Nobel complements and Nobel surrogates in the reward system of science. *Theoretical Medicine*, 13(2), 217-231.

Appendices

Appendix I: List of academic prizes by discipline

Table A.1: Academic prizes in economics and business

Economics and Business	Years considered	Award cycle
The Erwin Plein Nemmers prize in economics	2008-2016	biennial
Yrjö Jahnsson Award	2009-2017	biennial
Deutsche Bank Prize in Financial Economics	2007-2015	biennial
BBVA foundation frontiers of knowledge award in economics, finance and management	2012-2016	annual
IZA prize in labor economics	2012-2016	annual
The Stephen A. Ross prize in financial economics	2008-2016	biennial
Bernacer Prize	2012-2016	annual
Leontief Prize	2013-2017	annual
Global economy prize for economics	2013-2017	annual
The Ewing Marion Kauffman prize medal for distinguished research in entrepreneurship	2013-2017	annual

Table A.2: Academic prizes in life sciences

Life Sciences	Years considered	Award cycle
Crafoord prize in Biosciences	1999-2015	quadrennial
Darwin Medal	2008-2016	biennial
International Prize for Biology	2012-2016	annual
Louisa-Gross-Horwitz-Preis	2012-2016	annual
Heineken prize for biochemistry and biophysics	2008-2016	biennial
Breakthrough Prize in Life Sciences	2013-2017	annual
TWAS prize in Biology	2012-2016	annual
International cosmos prize	2012-2016	annual
ASBMB–Merck Award	2013-2017	annual
The Danone International Prize for Nutrition	2008-2016	biennial

Table A.3: Academic prizes in chemistry

Chemistry	Years considered	Award cycle
Wolf Prize in Chemistry	2013-2017	annual
Priestley Medal	2013-2017	annual
Welch award in chemistry	2012-2016	annual
NAS award in chemical sciences	2013-2017	annual
Faraday lectureship prize	2012-2016	annual
Davy medal	2012-2016	annual
Benjamin Franklin medal in chemistry	2013-2017	annual
Peter Debye award in physical chemistry	2013-2017	annual
Roger Adams award in organic chemistry	2009-2017	biennial
TWAS prize in chemistry	2012-2016	annual
Claude S. Hudson award in carbohydrate chemistry	2009-2017	biennial

Table A.3: Academic prizes in chemistry

Engineering	Years considered	Award cycle
Charles Stark Draper Prize	2012-2016	annual
John Fritz Medal	2012-2016	annual
Queen Elisabeth Prize for Engineering	2009-2017	biennial
Kyoto prize in advanced technology	2013-2017	annual
Kavli Prize in Nanoscience	2008-2016	biennial
Faraday Medal	2012-2016	annual
Millennium technology prize	2008-2016	biennial
TWAS prize in engineering sciences	2012-2016	annual
R.H. Wilhelm award in chemical reaction engineering	2012-2016	annual
Alpha Chi Sigma award for chemical engineering	2012-2016	annual
Founders award for outstanding contributions to the field of chemical engineering	2012-2016	annual
Andreas Acrivos Award for Professional Progress in Chemical Engineering	2012-2016	annual
Jacques Villermaux medal	1999-2015	quadren
Dieter Behrens medal	1997-2013	quadren
Freyssinet medal	2002-2014	quadren
International award of merit in structural engineering	2013-2017	annual
IABSE prize	2013-2017	annual
Theodore von Karman medal	2013-2017	annual
Fib medal of merit	2012-2016	annual
A.M. Turing Award	2012-2016	annual
IEEE medal of honor	2013-2017	annual
Benjamin Franklin medal in electrical engineering	2013-2017	annual
IEEE edison medal	2013-2017	annual
The Okawa prize	2012-2016	annual
The Knuth prize	2013-2017	annual
Royal Society Milner award	2013-2017	annual
Benjamin Franklin medal in computer and cognitive science	2013-2017	annual
W. Wallace McDowell award	2013-2017	annual
BBVA foundation frontiers of knowledge award in ICT	2012-2016	annual
World technology award in communications technology (for individuals)	2012-2016	annual
World technology award in it software (for individuals)	2012-2016	annual
World technology award in IT hardware (for individuals)	2012-2016	annual
Eni award	2012-2016	annual
The Enrico Fermi award	2010-2014	annual
The global energy prize	2012-2016	annual
World technology award in energy (for individuals)	2012-2016	annual
Tyler prize for environmental achievement	2013-2017	annual
Volvo environment prize	2012-2016	annual
Stockholm water prize	2012-2016	annual
BBVA foundation frontiers of knowledge award in ecology and conservation biology	2012-2016	annual
BBVA foundation frontiers of knowledge award in climate change	2012-2016	annual
Heineken prize for environmental sciences	2008-2016	biennial
The Zayed international prize for the environment	2008-2016	biennial
World technology award in environment (for individuals)	2012-2016	annual
Von Hippel award	2012-2016	annual
MRS medal award	2012-2016	annual
David Turnbull lectureship	2012-2016	annual
Materials Research Society: Outstanding Young Investigator Award	2012-2016	annual
World technology award in materials (for individuals)	2012-2016	annual
Royal society Armourers & Brasiers company prize	2008-2016	biennial
ASME medal	2013-2017	annual
Timoshenko medal	2013-2017	annual
Benjamin Franklin medal in mechanical engineering	2013-2017	annual
Gibbs brothers medal	2003-2017	triennial

Table A.4: Academic prizes in medicine

Medicine	Years considered	Award cycle
Albert Lasker Award for Basic Medical Research	2012-2016	annual
Lasker-DeBakey Clinical Medical Research Award	2012-2016	annual
Canada Gairdner international award	2013-2017	annual
Shaw Prize in Life Science and Medicine	2012-2016	annual
Canada Gairdner global health award	2013-2017	annual
Wolf Prize in Medicine	2013-2017	annual
Kavli Prize in Neuroscience	2008-2016	biennial
The Louis-Jeantet prize for medicine	2013-2017	annual
Robert Koch Preis	2013-2017	annual
Robert Koch Goldmedallie	2013-2017	annual
Lasker-Koshland special achievement award in medical science	2008-2016	biennial
King Faisal international prize for medicine	2013-2017	annual
Paul Ehrlich and Ludwig Darmstaedter prize	2013-2017	annual
Heineken prize for medicine	2008-2016	biennial
Lewis S. Rosenstiel Award	2012-2016	annual
Wiley prize in biomedical sciences	2013-2017	annual
Massry Prize	2012-2016	annual
Pearl Meister Greengard prize	2012-2016	annual
TWAS prize in Biology	2012-2016	annual
Crafoord prize in polyarthrits	2000-2017	quadren
J. Allyn Taylor international prize in medicine	2012-2016	annual
Jessie Stevenson Kovalenko Medal	2008-2016	biennial
Judson Daland prize for outstanding achievement in clinical investigation	2008-2014	varying
Tobias Prize	2008-2016	biennial
Albert Lasker Award for Basic Medical Research	2012-2016	annual

Appendix II: Survey details

Table A.5: Descriptive statistics

Variable	Unit	source	median	mean	s.d.	min.	max.
Research budget							
1st quartile	binary	Survey	0	0.25	0.43	0	1
2nd quartile	binary	Survey	0	0.25	0.43	0	1
3rd quartile	binary	Survey	0	0.26	0.44	0	1
4th quartile	binary	Survey	0	0.25	0.43	0	1
Academic rank							
junior	binary	Survey	0	0.04	0.20	0	1
postdoc	binary	Survey	0	0.26	0.44	0	1
assistant professor	binary	Survey	0	0.31	0.46	0	1
full professor	binary	Survey	0	0.38	0.49	0	1
Institution rank							
not ranked	binary	THE Ranking	0	0.36	0.48	0	1
tier 1	binary	THE Ranking	0	0.18	0.38	0	1
tier 2	binary	THE Ranking	0	0.23	0.42	0	1
tier 3	binary	THE Ranking	0	0.24	0.42	0	1
Controls							
age	count	Survey	45	46.21	10.85	25	88
female	binary	Survey	0	0.17	0.37	0	1
country							
Japan	binary	Survey	0	0.30	0.46	0	1
United Kingdom	binary	Survey	0	0.43	0.50	0	1
Germany	binary	Survey	0	0.27	0.45	0	1
scientific discipline							
biology	binary	Web of Science	0	0.27	0.44	0	1
chemistry	binary	Web of Science	0	0.31	0.46	0	1
economics	binary	Web of Science	0	0.20	0.40	0	1
engineering	binary	Web of Science	0	0.22	0.42	0	1
Publication information							
$\text{publications}_{2011\text{-}2016}$	count	Scopus	12	18.92	22.09	1	237
citations	count	Scopus	77	186.70	381.35	1	7332
citations per publication	fraction	Scopus	6	7.97	13.67	0	519
co-authors per publication	fraction	Scopus	5	5.42	2.98	1	45

Notes: Number of observations = 1884. Funding variables in million €, THE: Times Higher Education.

In the survey, the respondents were asked to answer research-related questions, especially about their employment situation, their institutional affiliations and their resource sharing behaviour. The survey provides several control variables that profile the respondents, including country, age, gender, academic position, and research budget. We exclude 23 individuals which appear as a principal investigator of the ERC project descriptions and two individuals which we identify to have won a prize from the sample.

The dataset contains survey responses from scientists in Germany (27%), Japan (30%), and UK (43%). We classified the respondents into four occupational ranks, i.e. junior scientists (4%), post-docs (26%), assistant professors (31%) and full professors (38%). The age of the respondents ranges between 25 and 88 years with an average age of 46 years. Among the respondents were 17% women. Using the provided annual research budget (with a median of 150.000€ and a mean of 4.7 million €), we create four budget categories, one for each quartile,

with each quartile containing nearly 25% of respondents. In order to control for institutional quality, we further lookup the institution rank of each respondent using international und country-ranking based in the "Times Higher Education Rankings". We classified the host institutions into a three-tier system (Tier 1: 0.18%; Tier 2: 23%; Tier 3: 24%) plus one class for those institutions that were not ranked (36%).

Appendix III: Document collection statistics

We refer to collections as sets of documents from a specific data source and in a specific field. In Table 4.2, we provide an overview of the twelve collections used in this study. These collections consist of four sample and eight frontier document collections. For each collection, we provide the number of authors, articles, total citations, and three ratios.

The sample comprises of 1884 authors with 27% in biology and medicine; 31% in chemistry, 20% in economics and business, and 22% in engineering. These authors have written 25842 articles in total and received more than 112577 citations during the years 2011-2016. The number of articles per author and the number of citation clearly varies between fields. Chemists have for example three times more articles per author than scholars in economics and business (29 vs. 8.8). We further see that biologists and chemists receive on average nearly twice as many citations per article, and clearly more citations per author, than economists and engineers.

Regarding academic prizes, we see that especially in chemistry and economics, there are fewer international awards than for biology and engineering (see Zheng and Liu 2015 for the prize population). The total 575 academic prize awardees have produced 14516 publications during 2011-2016. These 575 distinguished authors received nearly as many total citations as the total 2128 sample authors (381855 vs. 395407). In Biology, Chemistry and Engineering, scientists at the frontier of knowledge reveal a higher scientific productivity in terms of articles per author, compared to the sample authors. Only frontier economists seem on average to publish less but are nonetheless awarded with a science prize (5.9). Also the scientific impact, measured by citations, is clearly higher for frontier authors. Prize-winning engineers, for example, received more than 8 times the number of citations per author (101 vs. 823), while for economists this ratio is still twice as high.

Table A6: Collection statistics

		Biology & Medicine	**Chemistry**	**Economics & Business**	**Engineering**	**Total**
Sample authors	count	502	576	383	423	1884
Number of articles	count	9145	16697	3381	6416	25842
Total citations	count	90975	193934	21602	45236	112577
Articles per author	ratio	18.2 (16.0)	29.0 (29.4)	8.8 (10.2)	15.2 (18.8)	18.9 (22.1)
Citations per article	ratio	10.0 (24.0)	9.6 (7.1)	5.1 (6.2)	5.9 (5.1)	8.0 (13.7)
Citations per author	ratio	181.2 (238.3)	336.7 (588.0)	56.4 (147.1)	106.9 (195.8)	186.7 (381.4)
Academic prize awardees	count	214	53	52	256	575
Number of articles	count	4655	1828	306	7727	14516
Total citations	count	121976	44057	5074	210748	381855
Articles per author	ratio	21.7 (25.7)	34.5 (38.5)	5.9 (5.4)	30.2 (41.2)	25.6 (37.9)
Citations per article	ratio	26.2 (57.4)	24.1 (38.7)	16.6 (27.1)	27.3 (57.5)	26.3 (47.0)
Citations per author	ratio	570.0 (792.4)	831.3 (1122.8)	97.6 (143.4)	823.2 (1834.4)	672.3 (1458.8)
ERC project descriptions	count	1166	1369	509	1345	3114

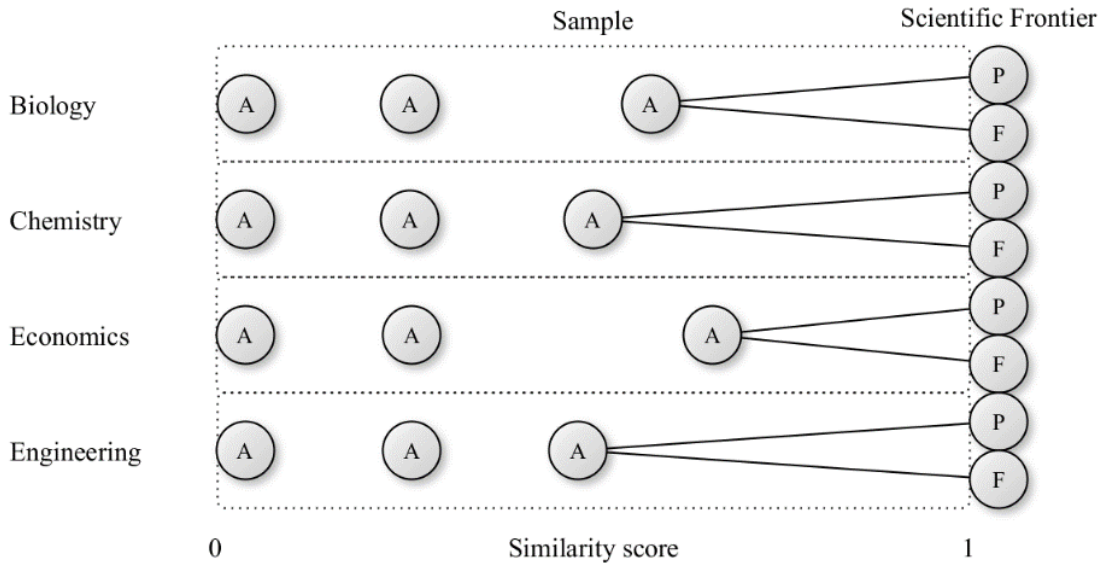
Notes: Publication records and project descriptions for the years 2011-2016, for the ratios we report the mean and standard deviation in parentheses

The project descriptions of the four ERC collections do not have bibliometric citations counts, which we can compare to the previous collections. However their magnitude reveals that economics and business projects are less often funded by the European Research Council.

Appendix IV: Illustration of similarity calculation

Figure A.1 illustrates the calculation of normalized similarity scores. For each author (A) in the sample we calculate a score that is indicative of how "close" he/she is on average to all knowledge frontier documents, either prizes (P) or funding (F), in the respective scientific field. A high score means that a sample author has on average more words in common with all frontier authors and thus seems to be "closer" to the frontier than a sample author with a lower score. This builds on the assumption that two authors work on a similar topic, if they share a common vocabulary.

Figure A.1: Document-document similarity between sample scientists and frontier science



Appendix V: Pre-processing and document representation

Rigorous pre-processing is crucial for co-word analysis and subsequent calculations. We start by removing punctuation, numbers and whitespaces from each document. Next, we remove generic terms (stop words) from each document using three "stop word" lists, i.e. the SMART list from the R package tm, a list of generic scientific vocabulary using the Academic Collocation List from Pearson Test of English Academic, and a custom list containing unusual fragments that we encounter during the analysis. We further use the stemming algorithm by Porter (1980) to truncating words to their word stem (Frakes and Baeza-Yates 1992).

We use the vector space model (Manning and Schütze 1999) to represents each document as a high dimensional vector where each dimension corresponds to a distinct term. A collection of m document vectors having a total of n terms will be represented by the $m \times n$ term-document-matrix A. Our goal is to transform A into a $m \times m$ similarity matrix S where $S_{i,j}$ gives some measure of the similarity between document vectors i and j. We specify four key parameters that condition the obtained similarity scores:

A. The first parameter refers to the unit of terms included in the matrix (token size). We include every single term (unigram) instead of a fixed term sequence such as bigram, trigram, n-gram etc. as a unit.
B. In addition, we remove extremely frequent words as well as seldom words that lie within specific bounds of the collection frequency (Frakes and Baeza-Yates 1992). We chose a lower bound of 3 and discard terms that occur in more than 33% of documents of the collection since we saw from the term frequency distribution that most terms above this threshold do not characterize a scientific specialization of authors.
C. The third parameter is the term length. While we do not set an upper limit here, we require the terms to have at least three characters to be meaningful.
D. We use Salton and Buckleys (1988) SMART weighting method to devalue non-discriminating terms while appreciating discriminating terms. Each term weight is the product of the term frequency, the collection frequency, and a vector length normalization scheme.

More specifically, we implement Salton and Buckleys augmented *tf/idf* weighting scheme with vector normalisation:

$$term\ weight_{t,d} = 0.5 + 0.5\ \frac{tf}{\max(tf)} \times \log_{10}(\frac{N}{\mathrm{n}}) \times\ \frac{1}{\sqrt{\sum tf^2}} \quad (2)$$

Table A.7: Calculation parameter overview

Parameter		Values	Description
A	token size	unigram, bigram	term sequence (unigram, bigram, trigram, n-gram)
B	collection bounds	3	minimum collection frequency (absolute)
		.33	maximum collection frequency (in percent)
C	term length	3	minimum number of characters in a term
		inf.	maximum number of characters in a term
D	term weight	augmented normalized tf	weight component for term frequency
		log(IDF)	weight component for inverse document frequency
		cosine normalization	weight component for document length

Appendix VI: Mapping of fields

Table A.8: Mapping of fields

Sample authors (ISA-Survey 2016)	Academic prize awardees (Scopus)	Prestigious research funding (ERC)
	Engineering	
Chemical, Thermal and Process Engineering, Computer Science, IT and Electrical and Electronic Engineering, Materials Science and Engineering, Mechanical, Aeronautical and Manufacturing Engineering, Civil and Construction Engineering; Architecture	Materials Science, Engineering, Energy, Computer Science, Chemical Engineering	Information Processing and Information Systems, Information and communication technology applications, Network technologies, Telecommunications, Electronics and Microelectronics, Physical sciences and engineering, Nanotechnology and Nanosciences, Space and satellite research, Aerospace Technology, Materials Technology, Industrial Manufacture, Construction Technology
	Economics/Business	
Arts and the Humanities, Business Administration, Economics	Arts and Humanities, Business, Management and Accounting, Economics, Econometrics and Finance, Decision Sciences, Social Sciences, Psychology	Social sciences and humanities, Business aspects, Economic Aspects, Regional Development
	Biology/Medicine	
Neurosciences, Agriculture, Forestry and Veterinary Medicine, Biological Sciences, Medicine (including Pharmacy, Dentistry and Nursing)	Neuroscience, Agricultural and Biological Sciences, Veterinary, Biochemistry, Genetics and Molecular, Biology, Immunology and Microbiology, Medicine, Pharmacology, Toxicology and Pharmaceutics, Nursing, Dentistry, Health Professions	Agricultural biotechnology, Life Sciences, Biotechnology, Medicine and Health, Medical biotechnology, Healthcare delivery/services
	Chemistry	
Chemistry, Geosciences (including Geography), Mathematics, Physics	Chemistry, Environmental Science, Earth and Planetary Sciences, Mathematics, Physics and Astronomy	Earth Sciences, Environmental Protection, Mathematics and Statistics, Physical sciences and engineering, Materials Technology

Appendix VII: Similarity measures

A large variety of measures have been proposed to express similarity, distance, or divergence between two statistical objects, e.g. tupel, vectors, sets or probability distributions (Lenz 2008; Deza and Deza 2009). These measures describe the statistical congruence between two document vectors that we wish to compare. The resulting similarity scores are usually high if two vectors have many common terms and low if not. While some coefficients are based on binary input, i.e. neglect the frequency with which a term occurs, others take into account the (weighted) frequency. The choice of similarity measure depends on the nature of data, the problem studied, and is not an exact science (Deza and Deza 2009).

Binary Similarity Models

Binary similarity measures do not use the term frequency directly and are rather based on four auxiliary variables (a-d). The binary models are defined by $t_{i,k} = 1$ if $tf_{i,k} > 0$ and $t_{i,k} = 0$ if $tf_{i,k} = 0$. Terms are thereof either present or absent in the document vectors. The auxiliary variables are defined as follows: For the i^{th} and j^{th} document, count $a_{i,j} = \sum_k t_{i,k} \times t_{j,k}$ is the number of mutual words present in both documents, $b_{i,j} = \sum_k t_{i,k} \times (1 - t_{j,k})$ and $c_{i,j} = \sum_k (1 - t_{i,k}) \times t_{j,k}$ represent words found in one document but not in the other. The number of terms that are mutually absent in both documents is denoted by $d_{i,j} = \sum_k (1 - t_{i,k}) \times (1 - t_{j,k})$ (see Table 4.4). Finally, n denotes the number of terms in the vectors. Using the auxiliary variables a-d, we implement the Jaccard Index, Sørensen–Dice Index, Simple Matching Coefficient, and Russel-Rao in our analysis (Table 4.5).

Table A.9: Auxiliary variables for binary similarity models

	Term present in Document 1	Term absent in Document 1
Term present in Document 2	$a_{i,j}$	$b_{i,j}$
Term present in Document 2	$c_{i,j}$	$d_{i,j}$

Metric Similarity Models

We further include four metric similarity measures that use the term frequency, i.e. cosine similarity, extended Jaccard, extended Dice and Pearson correlation. Cosine similarity is, considered as the "state of the art" in similarity measurement. In metric similarity measures, usually the dot product $\sum_k x_{i,k} x_{j,k}$ of the term weight is used in the numerator while different variants of normalization are used in the denominator.

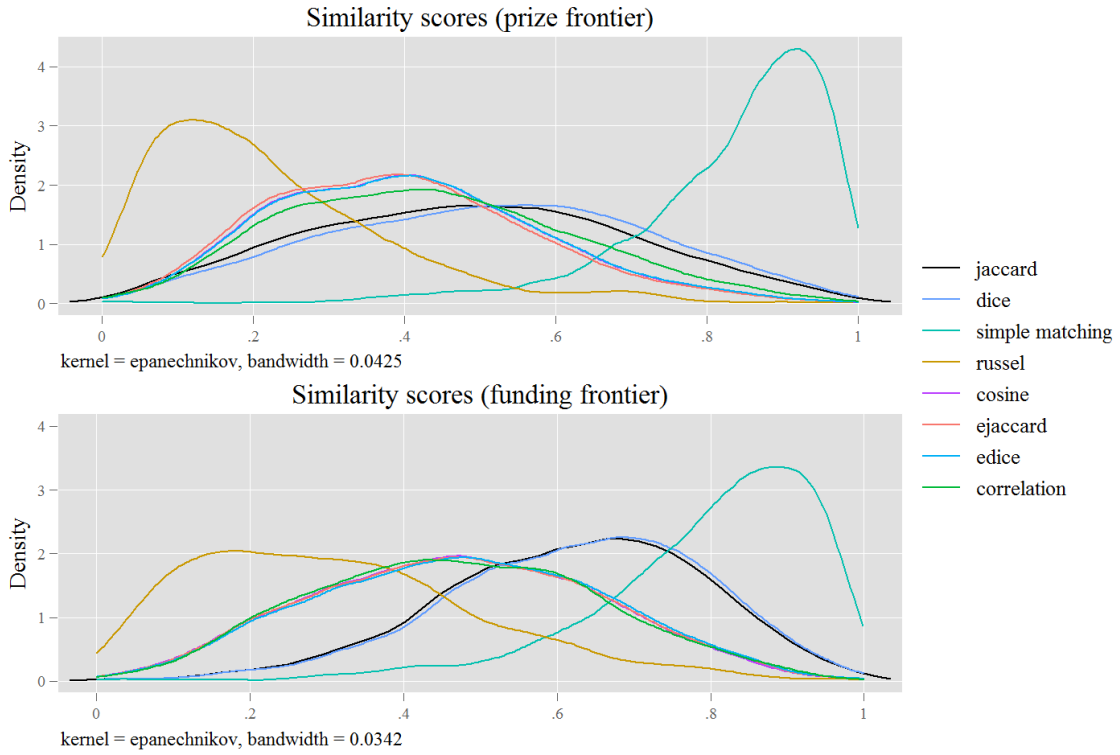
Table A.10: Similarity measure overview

	Similarity measure	Description	Formula	
Binary	Jaccard index	simplest index, size of the intersection divided by the size of the union, ignores d	$Sim_{i,j}^{(Jacc)} =$	$\frac{a_{i,j}}{(a_{i,j} + b_{i,j} + c_{i,j})}$
	Sørensen–Dice index	similar to Jaccard, greater weight to shared terms $a_{i,j}$	$Sim_{i,j}^{(Dice)} =$	$\frac{2a_{i,j}}{(2a_{i,j} + b_{i,j} + c_{i,j})}$
	Russel-Rao	intersection divided by total number of terms	$Sim_{i,j}^{(Russ)} =$	$\frac{a_{i,j}}{n}$
	Simple Matching Coefficient	similar to Jaccard index, takes terms into account that occur in none of the two documents	$Sim_{i,j}^{(SMC)} =$	$\frac{a_{i,j} + d_{i,j}}{n}$
Metric	Cosine similarity	state of the art, computes similarity as the vector normalized dot product of X and Y	$Sim_{i,j}^{(Cosi)} =$	$\frac{\sum_k x_{i,k} x_{j,k}}{(\sum_k x_{i,k}^2 \sum_k x_{j,k}^2)^{1/2}}$
	Extended Jaccard index	extension of the Jaccard index to metric data, equivalent to the binary version when the term vector entries are binary	$Sim_{i,j}^{(eJacc)} =$	$\frac{\sum_k x_{i,k} x_{j,k}}{(\sum_k x_{i,k}^2 + \sum_k x_{j,k}^2 - \sum_k x_{i,k} x_{j,k}}$
	Extended Sørensen–Dice index	extension of the Sørensen-Dice index to metric data	$Sim_{i,j}^{(eDice)} =$	$\frac{2\sum_k x_{i,k} x_{j,k}}{(\sum_k x_{i,k}^2 \sum_k x_{j,k}^2)}$
	Pearson Correlation	formally identical to the cosine similarity, invariant to shifts	$Sim_{i,j}^{(Corr)} =$	$\frac{\sum_k x_{i,k} x_{j,k}}{(\sum_k x_{i,k}^2 \sum_k x_{j,k}^2)^{1/2}}$ for centered weights

Appendix VIII: Average similarity scores

An overview of the normalized distribution for each similarity score gives Figure 4.3, where we plot the kernel density of eight similarity measures and both knowledge frontiers. In the case of the prize frontier (top of Figure 4.3), it becomes apparent that most average similarity scores follow a symmetric and flat normal distribution. This normality is confirmed by a visual test for normality using quantile-quantile-plots (not shown here). However there are two exceptions. The scores which are based on the Russel index are right-skewed (brown line), and those of the simple matching coefficient are left-skewed (green line), while both are more concentrated towards the lower and higher end of the score distribution. A deeper dive into their formulas in Table 4.5 reveals that this deviation is the result of incorporating "n", i.e. the total number of terms in the underlying highly sparse term-document matrix.

Figure A.2: Estimated distributions of normalized average similarity scores (N=1884)



A similar picture emerges for similarity score distributions using the funding frontier, but also for alternative parameter specifications.[8] While the density of the Russel index appears more normally distributed and less steep than in the case of academic prizes, the simple matching coefficient does not deviate much (lower part of Figure 4.3). The other two distributions based on binary measures (Jaccard, black line and Dice, blue line) peak in the third quantile and appear to be left-skewed. The four similarity scores based on metric measures do not differ much from those in academic prizes, however they are more symmetric and peak around the mode.

[8] To see how much these distributions depend on the parameter settings, we vary the token size between unigram and bigram, and the maximum term frequency between 33% (less restrictive) and 10% (more restrictive).