Detecting interdisciplinarity in top-class research using topic modeling

Andrea Bonaccorsi[1], Nicola Melluso[1] and Francesco Alessandro Massucci[2]

[1] *a.bonaccorsi@gmail.com, nicolamelluso@gmail.com*
Department of Energy, Systems, Territory and Construction Engineering, University of Pisa, Pisa (Italy)

[2] *francesco.massucci@sirisacademic.com*
SIRIS Lab, Research Division of SIRIS Academic, Barcelona (Spain)

Abstract

The paper applies topic modeling to the collection of ERC-funded proposals, interim reports and relative publications, with the aim of measuring in a novel way the degree of interdisciplinarity and addressing several open research questions which broadly aim at understanding how environmental conditions can favour the blossoming of interdisciplinarity. Without venturing into potential interpretations and explanations, we present a series of quantitative results linked with the above questions, while deliberately maintaining a descriptive attitude.

Introduction

The issue of interdisciplinarity (henceforth, ID) is again on top of the table for decision makers, funding agencies, and researchers as well. And, again and again, the issue remains open and problematic. Why is ID, at the same time, so requested and appreciated by society, and so difficult to adopt, implement, recognize, and prize by funding agencies, research institutions, and researchers?

In this paper we offer two contributions to a large and growing literature. On the technical side we experiment the use of advanced text mining techniques, namely optimal topic modeling, to describe and measure ID of research (henceforth, IDR). This follows from a recognition of the literature, that emphasizes the limitations of previous efforts to measure ID with the help of metadata of publications, such as classification schemes of journals (Subject Categories) and citations. Topic modeling has been applied to ID in a few papers in recent years, with promising results. We extend this line of investigation.

On the substantive side, we aim at addressing a number of research questions that are at the core of the debate on ID and which can be summarized as follows:

a) Do we see systematic differences in the level of ID by scientific discipline?

b) Is the propensity to undertake IDR influenced by organizational factors, such as the institutional affiliation (universities vs Public Research Organizations, henceforth PROs)?

c) Is the propensity to undertake IDR dependent, in a recognizable way, on the life cycle of researchers? Are young, or mid-career, or aged researchers more inclined to submit proposals with higher levels of ID?

These questions are clearly crucial in policy making: if decision makers want to promote IDR, for example to address complex societal challenges, they must know under which conditions IDR is likely to flourish.

To tackle the above questions, we examine the research projects funded by one of the world largest funding agencies, the European Research Council (ERC). From the open archive of ERC we extract the abstract of the funded research projects in the period 2007-2018 and integrate them with interim reports that funded projects must submit during the support period as well as with the abstract of the scientific publications produced by each project and indexed either in the open repository OpenAire or in the bibliometric database Scopus. For this documentation we extract the words that describe the research project (full text of the abstract) and the list of publications cited by each scholarly work linked to each project. We build a novel dataset by

reconstructing the biographic information of Principal Investigators (age, country, affiliation) and his/her list of publications (in particular, date of first publication).
We deliberately maintain an exploratory and descriptive attitude. There is large need to understand data qualitatively and with simple quantitative analysis, before entering into modeling exercises. We clearly have two limitations: due to institutional policies we can observe only winning project (so we are not in the position to run a counterfactual exercise), and submission to the ERC are not representative of the average research landscape but rather of the frontier, excellent research at the European scale.
The interest of the findings, however, will make justice of the proposed approach.

Related Works

In this section we address separately two streams of literature: the large stream of studies that try to define and measure ID (sub-sections 2.1 and 2.2) and the studies that identify possible antecedents at epistemic, institutional and individual level (sub-section 2.3).

Open issues in the definition and measurement of ID

ID has a complex meaning, that go deeply into epistemological and sociological debates. We do not enter here on these debates but follow the literature in assuming that ID is a multidimensional construct. Given the importance of the issue, there have been several efforts to identify, define and measure new indicators that might approximate the multidimensionality of the construct (Porter et al. 2006; Wagner et al. 2011). According to a classical conceptualization diversity can be defined in terms of (i) variety (number of disciplines); (ii) balance (distribution of disciplinary contributions); and (iii) disparity (distance or degree of difference between contributing disciplines) (Rao, 1982; Stirling, 2007). These definitions may lead to remarkably different measures of ID (Wang et al. 2015).
A consistent literature has defined ID by applying measures of diversity to the distribution of articles and citations by Subject Category (Morillo et al. 2001). According to this approach, the journals cited in the reference list of articles are classified by Subject Category and a citing-cited matrix is generated. After defining the metric of diversity it is possible to reduce the dimensionality of the resulting matrix and to project the disciplinary profiles on a map (Rafols et al. 2010; Carley and Porter, 2012; Cassi et al. 2017; Carley et al. 2017).
Following another approach, the diversity of disciplines is defined by the co-citation of articles in journals in different Subject Categories (Moya-Anegón et al. 2004; 2007; Porter et al. 2019). The most common definitions of diversity are the Rao-Stirling measures (Rao, 1982; Stirling, 1997), the Zhang et al. true diversity (Zhang et al. 2016), in addition to the standard measures of cosine distance (Porter et al. 2007; 2008) and Jaccard distance. Different measures of diversity place different emphasis to its dimensions and have different statistical properties, which are the object of critical examination. Klavans and Boyack (2009) have compared systematically the diversity measures and the resulting maps and support the use of cosine distance (see also Leydesdorff and Rafols, 2012). Ciotti et al. (2016) criticize the use of Jaccard measure as it is not independent on the size of the discipline. Fontana et al. (2018) support this view (i.e. the probability of integration between disciplines depends on the absolute and relative size of disciplines) and add a dynamic dimension (i.e. the probability of integration depends on the growth of disciplines). These contributions make it clear that the adoption of alternative definitions of ID and of diversity measures may lead to very different outcomes. As Leydesdorff and Rafols (2011) put it, "different indicators may capture different understandings of multifaceted concepts".

The Topic Modeling approach to ID

Partially in response to the limitations of the literature based on the structure of citations a recent stream of literature has been exploring the potential of Text mining techniques, applied to the words appearing in titles, abstract, or keywords of articles, less frequently on the full text. By using a variety of algorithms (in particular, Latent Dirichlet Allocation, LDA: Blei et al. 2002) and clustering procedures (Topic modeling) it is possible to represent the content of scientific papers (Rosen-Zvi et al., 2004; Lu and Wolfram, 2012; Gerrish and Blei, 2010; Wang et al. 2011; Nichols, 2014). In this case there is no need for ex ante classification, as in the Subject Category case, so that all limitations of the classification are removed. ID can be defined directly on the maps of topics.

The idea behind topic modeling as a technique for measuring ID has been taken on board by the National Science Foundation (henceforth, NSF) in the USA. According to NSF, the advantage of topic model approach on measuring ID is that its algorithm uses data drawn from the research itself, rather than from institutional structures through which the research was produced (Nichols, 2014).

Lu and Wolfram (2012) compare citation-based and word-based methods with topic-based techniques. By using LDA it is possible to uncover relationships between topics that would remain otherwise hidden due to terminological classifications. Also peripheral authors with few citations might be found central in terms of new topics. At the same time they call attention to the limits of topic modeling: there is no optimal rule for the determination of the number of topics, the label of the topic is assigned by convenience by the researcher, and the units of observation are single words and not more complex and long semantic units.

We emphasize that the methodological debate is not just of technical interest, but is crucial also for the hot debate on ID that takes place at epistemological level, as well as in research assessment and policy making. If measures of IDR are flawed, they deliver wrong messages to decision makers. There is therefore a keen interest in contributing to the methodology.

Why ID so difficult?

As stated above, ID is paradoxical, in the sense that there is large gap, so to say, between its demand, which is sustained and growing, and its supply, which is scarce. In this sub-section we explore three main areas that have been the object of disparate contributions, in which there are potential antecedents for this gap. They are: (a) epistemic, or disciplinary effects; (b) institutional effects; (c) individual, career-related effects.

Epistemic and disciplinary effects

Huutonieni et al. (2010) make the important distinction between addressing the issue from an instrumental and pragmatic perspective or from a more long term, conceptual, cognitive and epistemological perspective (Boon and van Baalen, 2019). From the latter perspective it must be admitted that we do not have a full scale theory of the intensity in which ID is practiced in different disciplines and of the underlying reasons. Several authors offered important insights, but no general results (Silva et al., 2013).

Somewhat less is known about ID in Social Sciences and Humanities. Garnier et al. (2013; 2014) evaluate a NSF program (Human and Social Dynamics) explicitly aimed at promoting ID within SSH and STEM (labeled cross-disciplinarity). They find a change in an integration measure before and after the program (2004-2008) and argue that without a specific program it would not have taken place spontaneously. They therefore recommend the adoption of targeted funding and specific requirements in grant calls.

On the other hand, researchers in SSH tend to cite contributions from STEM in an increasing way. Liu et al (2017) document a sharp increase in STEM citations in SSH papers in China but

warn that there is an inverse U-shaped relation between the extent of interdisciplinary citations and the scientific impact.
Summing up, we expect large differences across disciplines, but we do not have a theory supporting specific predictions.

Institutional effects

The literature has raised a number of concerns on the relation between institutional features of science systems and ID. Let us distinguish between funding institutions (government, funding agencies) and performing institutions (universities, PROs). With respect to funding, several authors have shown that IDR has consistently lower funding success (Metzger and Zare, 1999; Bromlsam et al. 2016; Kwon et al. 2017). Various explanations can be offered: panel of reviewers have less experience, the matching between projects and assessor expertise is more difficult, and it is easier to justify and explain the funding of conventional projects. Metzger and Zare (1999) strongly suggested the adoption of targeted funding programs.
With respect to performing institutions, much less is known. Universities may be at advantage since they have large variety of disciplines. In principle, social interaction among colleagues of different background may be facilitated (Cummings and Kiesler, 2005). Some studies have compared universities with a broad coverage (generalist universities) with dedicated or specialist universities (mainly in applied fields). Bonaccorsi and Secondi (2017) find that generalist universities outperform specialist ones in research productivity across all European countries (Bonaccorsi et al. 2021a; 2021b). In any case, these studies do not measure ID but overall productivity. On the other hand, when IDR requires prolonged interaction over time, PROs may be more flexible in arranging dedicated structures, relaxing researchers from the constraint of teaching. Also in this case we must explore the issue without a strong guidance from the literature. We are able to examine whether researchers who submit more interdisciplinary projects are affiliated to universities or PROs.

Age and career effects

An intriguing issue is the relation between the propensity of scientists to engage into IDR (either individually or within interdisciplinary teams) and their research career. We can distinguish two intertwined issues: a cognitive or epistemic issue, and an institutional or incentive issue. In some sense, they are the combination, at individual level, of the two powerful forces examined above, namely epistemic and institutional effects.
At the cognitive or epistemic issue we ask under what conditions do individual researchers conceive research ideas that require interdisciplinary knowledge to be addressed scientifically. Do they have deep issues in mind since their early years and develop research programs to address them later during their career? Or rather, do they focus initially on a narrow disciplinary focus and learn only later the need for opening their mind to other disciplines?
The cognitive or epistemic issue is inextricably linked to the problem of incentives. Is IDR good or evil for early stage researchers? Is the academic career enhanced or damaged by IDR?
Here the prevailing literature has been fairly sharp in proposing that the institutional systems based on publish or perish places severe penalties to IDR (Rafols et al. 2012; Chen et al. 2015; Wang et al. 2017). Other authors offer a more articulated view. For example, Lariviére and Gingras (2010) show that there is not a general pattern of correlation between ID and citations. There are large differences across disciplines. More recently, Okamura (2019) has challenged the negative view by showing that, in the population of highly cited papers (top 1% annual citation counts) increasing by one the number of disciplines leads to 20% increase in field normalized citations.
This debate is also important for policy makers: if they want more IDR but IDR is penalized in the academic career, simple financial schemes may not be enough.

Data and Methodology

Data

The analysis is conducted on the abstracts of the ERC scientific production. In particular, we focused on analysing the projects funded by the program from 2007 to 2018 and the publications which acknowledge financial support by those projects. This textual corpus is cross-linked with information on the respective researchers (i.e., to the PIs of the different projects) and with information about the institutions where those projects were effectively carried out. Since the funding scheme of ERC is conceived within an open access framework, we could make use of a series of open platforms to retrieve the above information: UNiCS[1], CORDIS[2], OpenAire[3] and ETER[4] (Mosca et al., 2018). These platforms give us the possibility to gather data about (i) the projects funded, (ii) the publications published thanks to the funding project, (iii) the principal investigators of the projects and (iv) the hosting institution of the project. Moreover, we gathered additional data at the level of publication, researcher and institution. With respect to publications, we extracted: the average reference age (namely the average distance between the publication year of a document and the publication year of the references) and the subject areas assigned by Scopus. For what concerns researchers, we collected their academic biographies (using Scopus) and we measured the distance (in year) from the winning year of the project to the year of their first publication (as PhD). Finally, for what concerns institutions, we gathered the foundation year (using ETER).

From topic models to ID

The measure of ID relies on the definition of an ID index calculated from the results of the topic modelling applied on the ERC scientific corpus. In particular, after removing the stop-words and stemming the texts, we used the LDA method to extract the topic models. As well as other unsupervised algorithms, LDA required a careful assessment of the reliability of the model resulted from the computation. In this case, we performed a concurrent analysis of 4 different metrics (Griffiths et al. 2004, Cao et al. 2009, Arun et al. 2010, Deveaud et al. 2014) to choose the best number of topics.

In LDA the topic distribution is assumed to have a sparse Dirichlet prior distribution (Ng et el., 2011). Considering the goal of the present paper, given a set of D documents and a chosen number of T topics, the main output of the LDA is a matrix of probability distribution (β_{dt}) that quantifies the probability β of a document d of belonging to the topic t. In this framing, an interdisciplinary topic would be one that groups documents spanning several pre-defined disciplinary classifications: for this reason, we shall consider the panel structure of the ERC program (which is structured into 3 main areas, namely Life Sciences, Physical Sciences and Engineering and Social Sciences and Humanities, which branch in turn into 25 panels) as the main reference point to calculate an index that explains the ID of a topic. When a principal investigator submits a project, she or he chooses a panel to which the contribution of the research belongs.

So that, given the Dirichlet distribution resulted from the topic models, we follow the assumptions of Leydesorff (2011): we determine the diversity of topics measuring how they are distributed among the three panels. Thus, the ID Index of a topic S_t is calculated as follows:

$$S_t = \frac{H(t)}{\log\ 3} = \frac{-(\gamma_{PE,t}\log(\gamma_{PE,t}) + \gamma_{SH,t}\log(\gamma_{SH,t}) + \gamma_{LS,t}\log(\gamma_{LS,t}))}{\log 3}$$

[1] https://unics.cloud/
[2] https://cordis.europa.eu/
[3] https://www.openaire.eu/
[4] https://www.eter-project.com/#/home

H(t) is the Shannon Entropy calculated among the probabilities that topic *t* belongs to the panels. PE, SH and LS represent respectively the "Physics and Engineering", "Social and Humanities" and "Life Sciences" panels.
Given the ID of a topic, it is possible to employ the document-topic probability matrix β_{dt} to calculate the ID of documents, researchers, institutions and disciplinary fields as follows:

$$S_d = \sum_{t=1}^{T} S_t \beta_{dt}; \qquad S_r = \frac{1}{n_r}\sum_{d=1}^{n_r} S_d; \qquad S_i = \frac{1}{n_i}\sum_{d=1}^{n_i} S_d; \qquad S_s = \frac{1}{n_s}\sum_{d=1}^{n_s} S_d;$$

S_r measures the ID a document d. S_r measures the ID of a researcher, with n_r equals the total number of documents produced by the researcher r. S_i measures the ID of an institution, with n_i the total number of documents produced by the institution i. S_s measures the ID of disciplinary field, with n_s the number of subject areas that Scopus attributes to the document. Note that S_r and S_i are calculated considering only the abstracts of the projects, while S_s is calculated considering only the publications.

Main findings

Descriptive statistics

We gathered a total sample of 41,379 abstracts (9,365 projects and 32,014 publications) and data about 1,062 institutions and 7,056 researchers. After performing the LDA algorithm, we evaluate the results identifying the best number of topics according to the assessment of state-of-the-art metrics (see crf. "*From topic models to ID*"). The best model resulted in 200 topics. The projects funded by the ERC and their related publications have an average degree of ID at around 0.45. The average degree grows over time. Figure 1 shows that the average degree is between 0.43 and 0.44 in 2008-2010 and 0.46 in 2018. We have no comparison with other institutions and/or periods, given the novelty of the indicator adopted. Nevertheless, the data show a slight increase in the decade if one discards the fluctuations of 2015-2016, which are generated by the change in Framework program between FP7 and H2020.

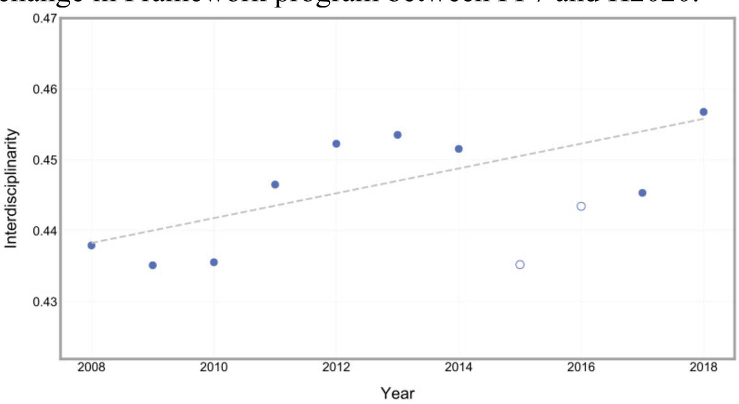

Figure 1. Dynamics of the ID degree. Year 2008-2018.

Disciplinary effect

There are large differences across disciplines in the degree of ID. The largest values are found in Physical and Engineering sciences (PE) in PE6 and PE10. On average, the degree of ID of PE fields is comparable to Life Sciences (LS).
Interestingly the lowest values are found in Social Sciences and Humanities (SH). The least interdisciplinary projects are found in SH2, SH5 and SH6, with SH1 at short distance.

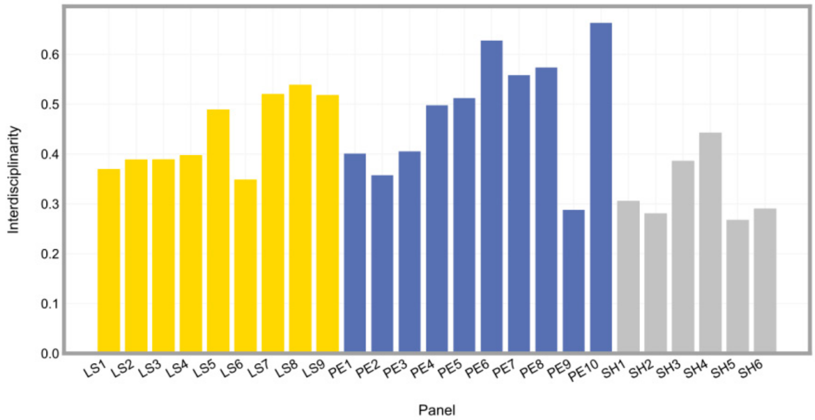

Figure 2. Degree of ID across disciplinary areas in the ERC panel.

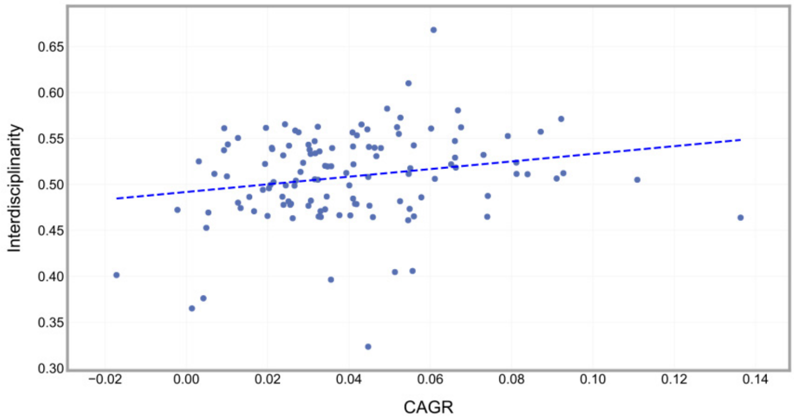

Figure 3. Relation between the rate of growth of publications (Compounded Annual Growth Rate 2008-2018) in the main field of the research project and the degree of ID.

There is a positive, although moderate, relation between the rate of growth of scientific disciplines and the degree of ID (Figure 3).

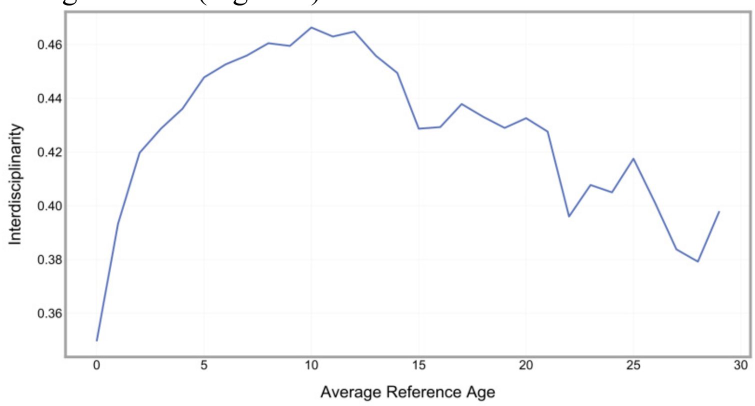

Figure 4. Relation between the average age of the papers cited in the ERC projects and the degree of ID.

We can interpret this finding in various ways. Fast growing disciplines may offer more opportunities for exploring the intersections and boundaries with other disciplines. Or, rather, they may result from the merge or convergence between previous established disciplines (as in the classical case of nanoscience), so that they maintain a dialogue with the originating fields. Or there is a demographic factor at place: fast growing fields may attract young researchers who are more oriented towards ID. For the time being, we cannot disentangle these effects.

On the contrary, we have a sharp result with respect to the structure of citations of other papers that the winners of ERC projects place in their documents. The list of publications cited in the abstract of the project can be considered a knowledge tree that makes explicit the origins of the research idea, as well as the state of the art that the applicants aim at innovating. What we find is that the projects with the largest degree of ID cite papers that are, on average, 10-11 year old (Figure 4). In other words, more interdisciplinary projects do not cite very old papers, but also do not cite very young papers. The former finding is perhaps trivial: applicants to ERC aim at improving over the state of the art and it would be difficult to believe that it includes very old pieces of knowledge. If they take as reference very old state of the art, it means that the field is not very dynamic (perhaps with the important, but very small, exception of unsolved mathematical problems, in the venerated tradition of Hilbert's list).
The latter finding is, on the contrary, surprising and interesting. Successful researchers that operate at the frontier do not use in their proposals and in the interim reports the very last papers. Is this in contradiction with the notion of frontier research? We offer an epistemic interpretation, as follows. ERC applicants do know they will be evaluated by top scientists in all disciplines and know that very young results need to be validated. The initial validation is likely to be done within disciplinary boundaries. There seems to be a *gestation period*, in which young results are progressively validated by their own discipline and become known outside the discipline. Remember that we compute the average age of the cited papers. The list of references may well include very young papers, but they are embedded into a knowledge structure which is relatively mature.

Institutional effects

It is perhaps true that PROs claim they support more IDR than traditional universities. The plain truth is that there is almost no difference between the two types of affiliation (Figure 5). No statistical test needed. [5]

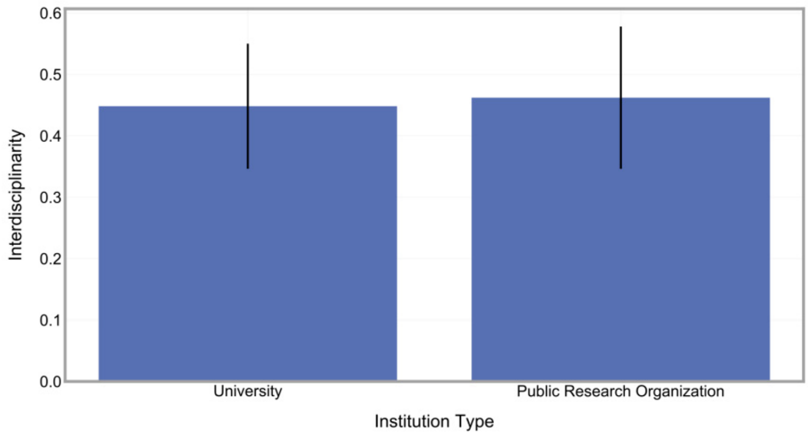

Figure 5. Average degree of ID by type of affiliation of ERC applicant.

In the absence of a robust theory that establishes a linkage between the institutional type (university vs PRO) and the propensity to engage into IDR, we leave this result to open discussion. On the contrary, we find moderate evidence that younger institutions are more likely to be the affiliation of ERC applicants with a higher degree of ID. With few exceptions, a high average degree of ID, as well as a larger variance across projects, is found in institutions established in the last two centuries (Figure 6). Combining the findings from Figure 5 and 6 it seems that the nature of the institution does not matter, but the age of the institution plays some

[5] Anyway, the ANOVA test of difference of means is rejected (p-value 0.278197).

role. We might be tempted to conclude that younger institutions try to differentiate from established institutions by adopting and nurturing a more interdisciplinary approach. The evidence however is only suggestive: conclusions are premature.

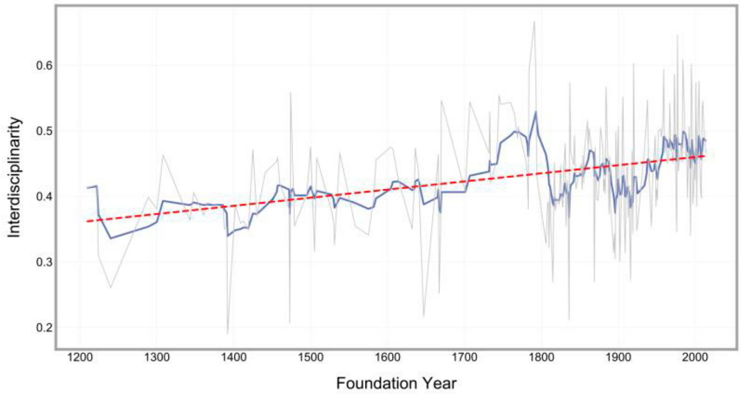

Figure 6. Relation between the age of the affiliation (foundation year of the institution) of ERC applicants and degree of ID.

Individual effects

As stated above, we have collected the biographies of the ERC winners and created a variable that measures the distance (in year) from the year of the PhD. This is an important information, since it dictates formally the eligibility of candidates to different ERC schemes. In particular, Starting Grants are dedicated to young researchers whose PhD degree has not been awarded more than 7 years before, while Advanced Grants do not place any constraint on the age of Principal Investigator. As it can be seen in Figure 8 the distribution of age of applicants has a modal value at 17 years after the PhD, while the bulk of the distribution is in the 10-17 years after the PhD. Researchers applying for Advanced Grants are not young post-doc, but rather young established scholars.

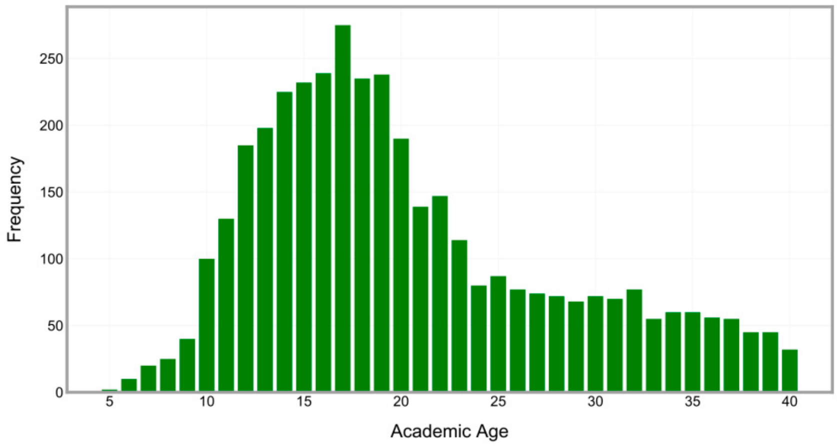

Figure 7. Distribution of ERC winners by academic age (number of years after the PhD).

Having such a large range in the distribution of academic age of ERC winners it is interesting to investigate whether more interdisciplinary projects are proposed by younger or older researchers. What we find is again striking: the peak of ID is achieved in projects in which the PI is 25 year older than the PhD, that is, is found in his/her ‘50s.

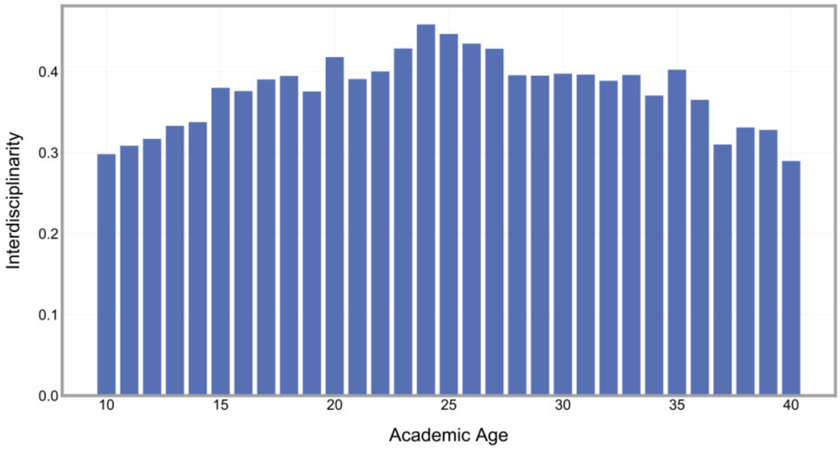

Figure 8. Distribution of the academic age of ERC applicants (number of years since PhD) by degree of ID.

Discussion and conclusions

The notion that IDR arises from young research and from young researchers is not supported. Young researchers must demonstrate first their contribution to their own discipline, focusing on internal topics, and react to the career and incentive systems. It seems that successful scientists that engage into highly interdisciplinary projects have already established their academic excellence. Whether this excellence has been gained with purely or interdisciplinary projects is not observable with our data and must be kept open. However, if ID would be an attribute of researchers, or a personal attitude, then we would see it at any age in the selected proposal. On the contrary, the degree of ID grows very little in the interval between 10 and 15 year of academic age.

An alternative conceptualization, which we find more persuasive, is a sort of life-cycle theory of individual propensity to ID. This conceptualization combines insights on the age of researchers and the age of cited literature. It goes as follows. All researchers are born disciplinary. Some of them maybe cultivate in their mind very audacious research questions that require out-of-the-box research approaches. But they do not take the risk of making them the core of their scientific production at the initial stages of the career. They know they are evaluated by senior colleagues on the basis of their contribution to the discipline. If they are successful they gain reputation and recognition. By going deeper and deeper into scientific problems they may come to appreciate contributions from other disciplines. They devote more time to read journals that are rarely cited by their colleagues but promising. They engage more freely into discussions without disciplinary boundaries. At some point in time they understand that entering into other fields, or collaborating actively with researchers from other fields, is promising, intellectually rewarding, creative, and original. They understand that the disciplinary recognition of this work will be delayed, because most of their colleagues will not accept or appreciate the break of boundaries. Nevertheless, they are no longer under the threat of delivering contributions that are immediately recognized by colleagues. They are established. Maybe their reputation may be at stake, not their tenure or salary. An intriguing question is how to verify whether researchers with a strong interdisciplinary orientation had this approach since their early years- but they did not show it in projects and publications. This would require additional research effort.

References

Arun, R., Suresh, V., Madhavan, C. V. & Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observation. In: *Pacific-Asia conference on knowledge discovery and data mining,* (pp. 391–402). Springer, Berlin, Heidelberg.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research,* 3, No. Jan, pp. 993-1022.
Bonaccorsi, A. & Secondi, L. (2017). The determinants of research performance in European universities: a large scale multilevel analysis. *Scientometrics*, 112(3), 1147-1178.
Bonaccorsi, A., Belingheri, P. & Secondi, L. (2021a). The research productivity of universities. A multilevel and multidisciplinary analysis on European institutions. *Journal of Informetrics*, 15, 101129.
Bonaccorsi, A., Belingheri, P. & Secondi, L. (2021b). Economies of scope between research and teaching in European universities. Submitted for publication.
Boon, M., & van Baalen, S. (2019). Epistemology for IDR. Shifting philosophical paradigms of science. *European Journal for Philosophy of Science*, 9:16.
Bromlsam, L., Dinnage, R. & Hua, X. (2016). IDR has consistently lower funding success. *Nature*, 534, 684-687.
Carley, S. & Porter, A.L. (2012). A forward diversity index. *Scientometrics*, 90(2), 407–427.
Carley, S., Porter, A.L., Rafols, I. & Leydesdorff, L. (2017). Visualization of disciplinary profiles. Enhanced science overlay maps. *Journal of Data and Information Science*, 2(3), 68-111.
Cassi, L., Champelmont, R., Mescheba, W., & de Turckheim, E. (2017). Analysing institution ID by extensive use of Rao-Stirling diversity index. *PLoS ONE*, 12(1), e0170296.
Cao, J., Xia, T., Li, J., Zhang, Y. & Tang, S. (2009). A density-based method for adaptive LDA model selection", Neurocomputing. 72(7–9), 1775–1781.
Chen, S., Arsenault, C. & Lariviére, V. (2015). Are top-cited papers more interdisciplinary? *Journal of Informetrics*, 9(4), 1034-1046.
Ciotti, V., Bonaventura, M., Nicosia, V., Panzarasa, P. & Latora, V. (2016). Homophily and missing links in citation networks. *EPJ Data Science*, 5:7.
Cummings, J.N.& Kiesler, S. (2005). Collaborative research across disciplinary and organizational boundaries. *Social Studies of Science*, 35(5), 703-722.
Deveaud, R., SanJuan, E. & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique,* 17(1), 61–84.
Fontana, M., Iori, M., Montobbio, F. & Sinatra (2018). A bridge over troubled water. Interdisciplinarity, novelty and impact. Università Cattolica del Sacro Cuore, Dipartimento di Politica Economica, *working paper* 2018/2.
Garnier, J., Porter, A.L., Borrego, M., Trau, E. & Tentonico, R. (2013). Facilitating social and natural science cross-disciplinarity. Assessing the human and social dynamics program. *Research Evaluation*, 22(2), 134-144.
Garnier, J., Porter, A.L. & Newman, N.C. (2014). Distance and velocity measures. Using citations to determine breadth and speed of research impact. *Scientometrics*, 100, 687-703.
Gerrish, S.M. & Blei, D.M. (2010). A language-based approach to measuring scholarly impact. ICML'10: *Proceedings of the 27th International Conference on Machine Learning*. June, 375-382.
Griffiths, T. L. & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(1), 5228–5235.
Huutonieni, K., Klein, J.T., Bruun, H. & Hukkinen, J. (2010) Analyzing ID. Typology and indicators. *Research Policy*, 39, 79-88.
Klavans, R. & Boyack, K.W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology*, 60(3), 455-476.
Kwon, S., Solomon, G.E.A., Youtie, J. & Porter, A.L. (2017) A measure of knowledge flow between specific fields. Implications of ID for impact and funding. *PLoS ONE*, 12(10), e0185583.
Lariviére, V. & Gingras, Y. (2010). On the relationship between ID and scientific impact. *Journal of the American Society for Information Science and Technology*, 61(1), 126-131.
Leydesdorff, L. (2007). Betweenness centrality as an indicator of the ID of scientific journals. *Journal of the American Society for Information Science and Technology*, 58(9), 1303-1309.
Leydesdorff, L. & Rafols, I. (2011). Indicators of the ID of journals. Diversity, centrality, and citations. *Journal of Informetrics*, 5, 87-100.
Leydesorff, L. & Rafols, I. (2012). Interactive overlays. A new method for generating global journal maps from Web of Science data. *Journal of Informetrics*, 6, 318-332.

Liu, M., Shi, D. & Li, J. (2017). Double-edged sword of interdisciplinary knowledge flow from hard sciences to humanities and social sciences. Evidence from China. *PLoS ONE*, 12(9), e0184977.
Lu, K. & Wolfram, D. (2012). Measuring author research relatedness. A comparison of word-based, topic-based and author cocitation approaches. *Journal of the American Society for Information Science and Technology*, 63(10), 1973-1986.
Metzger, N. & Zare, R.N. (1999). IDR. From belief to reality. *Science*, 283(5402), 642-643.
Morillo, F., Bordons, M. & Gómez, I. (2001). An approach to ID through bibliometric indicators. *Scientometrics*, 51(1), 203-222.
Moya-Anegón, F., Vargas-Quesada, B., Chinchilla-Rodriguez, Z., Corera-Álvarez, E., Munoz-Fernández, F.J. & Herrero-Solana, V. (2007). Visualizing the marrow of science. *Journal of the American Society for Information Science and Technology*, 58(14), 2167-2179.
Moya-Anegón, F., Vargas-Quesada, B., Herrero-Solana, V., Chinchilla-Rodriguez, Z., Corera-Álvarez, E. & Munoz-Fernández, F.J. (2004). A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics*, 61(1), 129-145.
Mosca, A., Roda, F. & Rull, G. (2018). UNiCS-The Ontology for Research and Innovation Policy Making. *Frontiers in Artificial Intelligence and Applications*, 306, 200-207.
Nichols, L.G. (2014). A topic model approach to measuring ID at the National Science Foundation. *Scientometrics*, 100, 741-754.
Porter, A.L., Carley, S., Cassidy, C.N., Youtie, J., Schoeneck, D.J., Kwon, S. & Salomon, G.E.A. (2019). Measuring IDR categories and knowledge transfer. A case study of connections between cognitive science and education. *Perspectives on Science*, 27(4), 582-618.
Porter, A.L., Cohen, A.S., Roessner, J.D. & Perreault, M. (2007) Measuring researcher ID. *Scientometrics*, 72(1), 117-147.
Porter, A.L. & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, 81(3), 719-745.
Porter, A.L., Roessner, J.D., Cohen, A.S. & Perreault, M. (2006). IDR: Meaning, metrics and nurture. *Research Evaluation*, 15(3), 187-195.
Porter, A.L., Roessner, J.D. & Heberger, A.E. (2008). How interdisciplinary is a given body of research? *Research Evaluation*, 17(4), 273-282.
Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P. & Stirling, A. (2012). How journal rankings can suppress IDR. A comparison between innovation studies and Business & Management. *Research Policy*, 41, 1262-1282.
Rafols, I., Porter, A.L.& Leydesdorff, L. (2010). Science overlay maps. A new tool for research policy and library management. *Journal of the American Society for Information Science and Technology*, 61(9), 1871-1887.
Rao, C.R. (1982). Diversity. Its measurement, decomposition, apportionement and analysis. *Sankhya: The Indian Journal of Statistics*, 44(1), 1-22.
Rosen-Zvi, M., Griffiths, T., Steyvers, M. & Smith, P. (2004). The Author-topic model for authors and documents. *Proceedings of the UAAI Conference*, 487–494.
Silva, F.N., Rodriguez, F.A., Oliveira, O.N. & da Costa, L. (2013). Quantifying the ID of scientific journals and fields. *Journal of Informetrics*, 7, 469-477.
Stirling, A. (1997). On the economics and the analysis of diversity. *SPRU Working Paper*, no. 28.
Wagner, C., Roessner, J., Bobb, K., Klein, J., Boyack, K., Keyton, J., et al. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics*, 5(1), 14–26.
Wang, H., Ding, Y., Tang, J., Dong, X., He, B., Qiu, J., et al. (2011). Finding complex biological relationships in recent PubMed articles using Bio-LDA. *PLoS One*, 6(3), 1–14.
Wang, J., Thijs, B. & Glänzel, W. (2015). Interdisciplinarity and impact: Distinct effects of variety, balance, and disparity. *PLoS ONE*, 10(5): e0127298
Wang, J., Veugelers, R. & Stephan, P. (2017). Bias against novelty in science. A cautionary tale for users of bibliometric indicators. *Research Policy*, 46, 1416-1436.
Zhang, L., Rousseau, R. & Glänzel, W. (2016). Diversity of references as an indicator for ID of journals. Taking similarity between subject fields into account. *Journal of the American Society for Information Science and Technology,* 67(5), 1257-1265.