TÉCNICO LISBOA

How to measure productivity, quality, collaboration and innovation in Health Networks?

A bibliometric study of the University of Lisbon

Maria Carolina Manique Bento Florindo da Conceição

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisors: Dr. Paulo Jorge de Morais Zamith Nicola
Prof. Mónica Duarte Correia de Oliveira

Examination Committee

Chairperson: Prof. João Miguel Raposo Sanches
Supervisor: Dr. Paulo Jorge de Morais Zamith Nicola
Member of the Committee: Prof. Fernando Fernandez-Llimos Somoza

November 2021

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Preface

The work presented in this thesis was developed during the period of March 2021 to November 2021 at the Unidade de Epidemiologia (UEPID) in the Faculdade de Medicina da Universidade de Lisboa, FMUL, under the supervision of Dr. Paulo Nicola. This thesis was co-supervised at Instituto Superior Técnico by Prof. Mónica Oliveira.

Acknowledgments

Ao professor Paulo Nicola, agradeço a presença diária incansável, o constante bom humor, optimismo e encorajamento perante as adversidades. Agradeço também à professora Mónica Oliveira os conselhos e palavras sábias e a compreensão e preocupação sempre demonstradas.

Obrigada, mãe e pai, por me terem ajudado a chegar até aqui. Obrigada por me terem dado tudo o que deram e por me terem incentivado a querer sempre mais para mim.

Aos meus amigos, obrigada pelo apoio durante estes duros meses, pelas vezes que me puxaram para cima, pelas inúmeras chamadas de estudo, por todos os risos e choro e obrigada, acima de tudo, por sempre terem acreditado em mim. Um agradecimento especial à minha amiga Rita Silva, que desde o primeiro ano que está pronta para ouvir e responder às minhas questões e esta tese não foi exceção.

Abstract

Scientific research plays a fundamental role in the development of society. Bibliometrics, the use of analytical and mathematical methods to study publications, has been developed to examine this research. The purpose of this thesis is to characterize the scientific production of the University of Lisbon, between 2014 and 2019, mainly in the health domain.

To accomplish this goal, the data of the publications of the institution were retrieved from the Web of Science. These data were then standardized and publications from the health domain were selected. After that, publication and citation counts, co-authorship relationships and keywords co-occurrences networks were obtained.

Production in the health domain grew in the periods in study from 4102 in 2014 and 2015 to 5947 in 2018 and 2019. The impact of publications of the health domain, as measured by the average number of citations per document per year, also increased regarding the first time period. Collaboration, as measured through co-authorship, has increased. Finally, innovation, as measured through the number of unique keywords and using the keywords' co-occurrence network, also presented a growing trend.

The scientific production of the University of Lisbon seems to be in ascension. More bibliometric studies should be performed to analyse this evolution, as well as to inform policy making.

Keywords

bibliometrics; science mapping; measuring research; social networks analysis; collaboration

Resumo

A investigação científica desempenha um papel fundamental no desenvolvimento da sociedade. A bibliometria, isto é, a utilização de métodos matemáticos e estatísticos para estudar publicações científicas, foi criada para analisar esta investigação. O objetivo desta tese é fazer a caracterização da produção científica da Universidade de Lisboa, entre 2014 e 2019, particularmente no domínio da saúde.

Os dados utilizados neste estudo foram obtidos através do Web of Science. Estes dados foram posteriormente padronizados e as publicações do domínio da saúde foram selecionadas. A partir da informação obtida, foram realizadas contagens de publicações e citações e foram identificadas redes de colaboração entre instituições e redes de co-ocorrência de palavras chave.

A produção, no domínio da saúde, aumentou de 4102 em 2014 e 2015 para 5947 em 2018 e 2019. O impacto das publicações deste domínio, dado pela média do número de citações por documento por ano desde a publicação, também aumentou em relação ao primeiro período em estudo. A colaboração, medida através da coãutoria de publicações, também apresentou sinais de crescimento. Finalmente, a inovação, medida através do número de palavras chave únicas, bem como atravás rede de co-ocorrência de palavras chave, apresentou um tendência ascendente.

A produção científica na Universidade de Lisboa parece estar em ascenção. Neste sentido, deve apoiar-se a aplicação de mais estudos bibliométricos para analisar esta evolução e para informar a tomada de decisões.

Palavras Chave

bibliometria; mapeamento das publicações científicas; análise de redes sociais; colaboração

Contents

List of Figures

List of Tables

Acronyms

DB	Database
DOI	Digital Object Identifier
EU	European Union
FAUL	Faculdade de Arquitetura da Universidade de Lisboa
FBAUL	Faculdade de Belas Artes da Universidade de Lisboa
FCS	Field Citation Score
FCUL	Faculdade de Ciências da Universidade de Lisboa
FDUL	Faculdade de Direito da Universidade de Lisboa
FFUL	Faculdade de Farmácia da Universidade de Lisboa
FLUL	Faculdade de Letras da Universidade de Lisboa
FMD	Faculdade de Medicina Dentária da Universidade de Lisboa
FMH	Faculdade de Motricidade Humana
FMUL	Faculdade de Medicina da Universidade de Lisboa
FMV	Faculdade de Medicina Veterinária da Universidade de Lisboa
FPUL	Faculdade de Psicologia da Universidade de Lisboa
IDL	Instituto Dom Luiz
IE	Instituto de Educação
IGOT	Instituto de Geografia e Ordenamento do Território
iMM	Instituto de Medicina Molecular
INESC-ID	Instituto de Engenharia de Sistemas e Computadores, Investigação e Desenvolvimento
ISA	Instituto Superior de Agronomia
ISC	Instituto de Ciências Sociais
ISCSP	Instituto Superior de Ciências Sociais e Políticas

ISEG	Instituto Superior de Economia e Gestão
IST	Instituto Superior Técnico
JIF	Journal Impact Factor
MUHNAC	Museu Nacional de História Natural e da Ciência
NLM	National Library of Medicine
OE	Organization Enhanced
SCI	Science Citation Index
SNA	Social Network Analysis
TUL	Technical University of Lisbon
UL	University of Lisbon
USA	United States of America
WHO	World Health Organization
WoS	Web of Science

1

Introduction

Contents

1.1 Context

The human era has always been characterized by progress and development. From the discovery of fire more than a million and a half years ago, followed by the invention of the wheel six millennia ago, these advancements have promoted the improvement of the quality of life for the vast majority of human beings. Food production has evolved, diseases have been eliminated or cures have been found and scientific research has contributed to the brightening of the world the human being lives in.

In the eighteenth century, universities, as they are known today, started to emerge. These institutions where both the activities of teaching, or learning, and research were performed and walked hand in hand surged. After the second World War, due to the influence of the scientific progress promoted by universities, agriculture and manufacturing became increasingly efficient having to depend less and less on human-labour. There was, then, a shift from the primary and secondary sector of activity to the services industry, which required a higher degree of specialization of workers [3]. The crescent number of students enrolling in universities boosted the development of these institutions, which nowadays contribute as much to innovation, the motor of progress, as governments or firms.

This rapid advancement of scientific research and of the academia brought new questions: how is scientific research impacting our lives? What topics are being studied? How can scientific research be improved? Well, luckily, nowadays there are thousands of tools that can be used to find an answer to these questions. The crescent computational power at the disposal of the human mind can be used to create an unprecedented understanding of science. Thus, the emergence of the multidisciplinary field "Science of Science". Quantifying or measuring scientific production, impact, collaboration and innovation may enable the discovery of unrivaled potential in all scientific fields, thus accelerating progress and the improvement of this amazing world [4].

Progress was not constructed alone, though. As science evolved, so did the way in which it was done. In the past century, the number of authors, teams, institutions, even countries, who collaborated in a single publication has been increasing [4–6]. Collaboration has been shown to increase productivity [7], impact [7, 8], funding acquisition [4, 5, 9] and even innovation [4]. However, this increase in collaboration and, consequently, its advantages are becoming more and more concentrated in the most prestigious institutions [6]. There is an homophily context in which the highest ranking institutions end up collaborating with each other and benefiting the most of this new reality contributing to crescent disparities between institutions [4]. It is therefore of extreme importance for a university to be familiar with its collaboration network and understand how to improve it as not to be left behind in the pursuit of science.

A powerful ally in the pursuit of characterizing science is the field of bibliometrics. Bibliometrics consists in the application of statistical and mathematical methods to assess several characteristics of scientific research. The first time someone performed a systematic acquisition of bibliographic data was in 1906 when Cattell, an American scientist and editor of *Science* between 1895 to 1944, launched the

American Men of Science, which gathered information about influential scientists in the United States of America (USA) [10]. Several scientometric and bibliometric studies have been performed since then, but only with the creation of the Science Citation Index (SCI) did this domain see a surge in its development. Nowadays, bibliometrics is widely used to quantify and characterize the scientific research of individual researchers, research groups, institutions, journals and even countries. These data and analysis should not be the sole information in which research managers or policy makers base their decisions in. However, allied with other methods, such as peer review, and a profound self-awareness of its own limitations, this field can provide the information needed to make reforms and improve the scientific endeavor.

When searching for publications on bibliometrics with a Portuguese address in the Web of Science (WoS), 112 publications can be retrieved. A quick read of the titles of those publications makes it clear that bibliometric studies have not yet been applied to individual Portuguese institutions or departments, not even research groups. This may reveal that research on the field of bibliometrics, in Portugal, is not acquiring enough interest from researchers, institutions or even funding agencies. In 2008, the use of bibliometric indicators was discouraged by the FCT (Fundação para a Ciência e Tecnologia) in the assessments made by peers [11]. However, as noted by Ramos and Serrica [11], "a significant number of references to bibliometrics can be found in the evaluation reports of ALabs" and those references were used in arguable ways, "the panels resorted to a non-homogeneous set of 'indicators', including the questionable practice of using journal rankings or impact factor as proxies of research impact". Even though that is not the focus of this thesis, it seems relevant to highlight that the field of research policy in Portugal may not be taking full advantage of bibliometrics. For example, in the eighties, Moed *et al.* devised a large scale study in order to inform funding allocations in the University of Leiden [12].

Additionally, bibliometrics plays an even more prominent role in health research. This area takes advantage of developments in so many other domains that more and more data, more and more publications should be accounted for. Bibliometrics makes it easier for fields to be mapped, evolution traced and breakthroughs analysed, thus accelerating progress in the health domain [13–15]. This thesis aims to contribute to bibliometric literature in Portugal through the development and application of methods that characterize scientific production and collaboration of the publications of the University of Lisbon (UL), specially in the health domain.

The UL was legally founded by the Portuguese government in the decree of law of the 19th of April 1911. The Technical University of Lisbon (TUL) was founded in 1930 by the decree of law of the 2nd of December of 1930. In 2013, both universities were merged to create the biggest university in Portugal and the fourth most sizable in the Iberian Peninsula. Nowadays, it counts with 18 schools (Faculdade de Arquitetura da Universidade de Lisboa (FAUL), Faculdade de Belas Artes da Universidade de Lisboa (FBAUL), Faculdade de Ciências da Universidade de Lisboa (FCUL), Faculdade de Direito da Uni-

versidade de Lisboa (FDUL), Faculdade de Farmácia da Universidade de Lisboa (FFUL), Faculdade de Letras da Universidade de Lisboa (FLUL), Faculdade de Medicina da Universidade de Lisboa (FMUL), Faculdade de Medicina Dentária da Universidade de Lisboa (FMD), Faculdade de Medicina Veterinária da Universidade de Lisboa (FMV), Faculdade de Motricidade Humana (FMH), Faculdade de Psicologia da Universidade de Lisboa (FPUL), **ICS!** (**ICS!**), Instituto de Educação (IE), Instituto de Geografia e Ordenamento do Território (IGOT), Instituto Superior de Agronomia (ISA), Instituto Superior de Economia e Gestão (ISEG), Instituto Superior de Ciências Sociais e Políticas (ISCSP), Instituto Superior Técnico (IST)) and more than 100 research units [16].

The bibliometric indicators that are available in the webpage of the UL are the number of publications and the number of citations, referent to the years of 2013 to 2015 [17]. Besides that, a study of the publications in the health domain of the UL has not been performed yet, to the best of the knowledge of this author and advisors. Having in mind the existence of RedeSaúde UL, an analysis of such publications is even more relevant. RedeSaúde UL constitutes a network of the UL, whose goal is to organize researchers and find opportunities for collaboration to provide an harmonized front to deal with the challenges the health research field presents today [18]. Having that in mind, a description of both production and collaboration in this domain in the UL is important to quantify, and, consequently, understand, the productivity of this institution in this domain, as well as the subdomains in which research is produced, the extension of collaboration and innovation generated by this institution, so that plans for improvement can be put into place where they are needed and to explore the full potential that the UL has to offer.

1.2 Objectives

Taking all the factors described in the previous section into account, this thesis aims to develop and implement methods to describe the scientific production of the University of Lisbon (UL) in all domains and in the health domain in particular, as well as to measure collaboration and innovation potential of the university in this domain.

1.3 Thesis Outline

This thesis is organized as follows: a literature review chapter (chapter 1) is followed by the methodology (section 2.3), after that the results obtained from the implementation of the methodology are presented (chapter 4) and are then discussed (chapter 5), finally there is a chapter presenting the main conclusions of this study (chapter 6).

Firstly, in the literature review, the concepts important for the fulfillment of this thesis are presented

based on previous literature on the area. This chapter is split in a review of bibliometrics, corresponding tools and bibliographic databases, followed by a review of collaboration and the methods used to understand and measure it, followed by a section on innovation.

After that, the proposed methodology developed on this thesis, that is grounded on previous bibliometric methods and studies performed and exposed in the literature review, used to obtain the results is introduced. This chapter starts with an introduction on the research questions and selection of methods to perform the study, followed by the approach used to retrieve and clean or standardize the data. Then, the metrics used to perform the analysis are introduced. Finally, the approaches used to visualize the data are presented.

The results presented in chapter 4 are reported in line with the proposed methodology. In this chapter, the productivity, content, innovation, impact and collaboration of the UL's scientific production overall and for the health domain are described. Finally, the collaboration and keyword co-occurrence's networks are shown and analysed.

Afterwards, the results are discussed. Besides that, there is the presentation of the benefits and shortcomings of the methodology used in this work.

Finally, this thesis is concluded with a brief summary of the results and of the discussion of those same results. Additionally, recommendations and future perspectives for this area of work are proposed.

2

Literature Review

Contents

The aim of this thesis, as stated in chapter 1, is to describe the scientific production, as well as the impact, of the research performed in the UL, specifically in the Health domain. To accomplish those goals bibliometric methods and Social Network Analysis (SNA) will be used. Therefore in this section, these concepts are going to be introduced along with how they can be used to assess research productivity and impact, collaboration and innovation.

Most of the concepts mentioned in the previous paragraph are concepts that do not have a substantial expression in the biomedical curriculum and as such the literature review was of grand importance. The literature review was initially performed by searching for keywords, on Google Scholar. The choice of keywords can be found in table 2.1. And then after finding the relevant work, cycling was performed. "Cycling means examining the bibliographies of the papers you start with, and of the source papers obtained in order to locate additional relevant work" [19]. Inverse cycling was also performed using the Google Scholar tool for citing papers.

Table 2.1: Keywords used to retrieve documents relevant to the literature review in the three different areas, bibliometrics, collaboration and innovation.

Field	Keywords
Bibliometrics	Bibliometrics Research Performance Research Evaluation Science Mapping
Collaboration	Collaboration Collaboration Network Research Collaboration Social Network Analysis
Innovation	Innovation Scientific Research Innovation Measuring Innovation

This chapter's first section focuses on literature on bibliometrics and its methods. This section starts with a brief introduction of the concept, followed by a review of the indicators, as well as its uses and possible limitations, after that science mapping is explored and lastly the limitations of bibliometrics are discussed. From there on, a review of literature on collaboration and the means to measure it is presented, which starts with an introduction to scientific collaboration, an explanation about how co-authorship can be used to study collaboration, as well as SNA. Finally, the last section deals with innovation and the tools to analyse it.

Nowadays, bibliometrics is contained within the field of informetrics [20]. However, this field was the first to historically arise [21]. Bibliometrics started to be referred to as "statistical bibliography" and the term bibliometrics seems to have been first used by Otlet, a french researcher, but it was Pritchard that ended up coining the name and defining bibliometrics [20, 21] and it can be thought of as the measure through statistical and mathematical methods of scientific literature. Scientometrics is the quantitative

study of many aspects of science, such as its economy, and as such it may inform the decision making of science policy makers. It may also rely on the study of publications and due to that, it crosses bibliometrics' boundaries in certain situations. Webometrics constitutes the study of written documents that are present on the web. As it is text based, it is totally embraced by bibliometrics. Finally, informetrics goes beyond the study of bibliographic data or scholarly communication and thus it contains all the other fields [20, 21]. A schematic organization of these fields can be observed in figure 2.1.

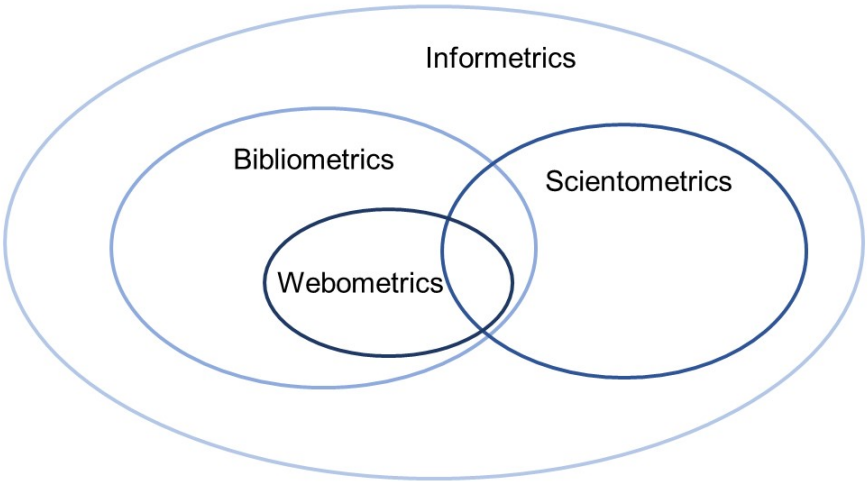

Figure 2.1: Diagram representing the overlapping between the different areas of informetrics. This diagram was adapted from Thelwall *et al.* [1].

2.1 Bibliometrics

According to Cavadas, "bibliometrics summarises a set of quantitative and statistical methods applied to analyse publication patterns, allowing to obtain insights from a macro perspective on the structure and collaborative relationships of scientific activities" [22]. The use of bibliometrics for the evaluation of scientific production as well as to inform research policy and funding allocation all over the world has been increasing. These methods can be regarded as an objective means to mitigate the biases that may arise in the peer review process and as such should be used in parallel with peer evaluation to ensure consistency and impartiality [23, 24]. Besides this, bibliometrics can also be used improve the decision-making process of individual researchers, by assessing mechanisms through which they can increase their impact [25] and it can also be used to map the cognitive links and identify sub-domains inside a field of study or between fields of study [23, 26], therefore aiding policy-makers in making decisions regarding

research management. However, how can these methods be used to aid in these endeavors? By providing quantitative and objective measures of the productivity and impact of a researcher, institution, journal or country.

2.1.1 Bibliometric indicators

As stated in the previous paragraph, in order to make use of bibliometrics to assess scientific production, there are some fundamental concepts and assumptions that one has to have in mind. Firstly, the number of publications is a proxy for scientific production. Besides that, the number of citations is a proxy for the impact of a publication. Finally, co-publication is a proxy for collaboration. The first idea is based on the fact that, especially in the natural sciences and in the health domain, a publication is both a vector for information and a way to coin an idea as one's own and as such it is a means to convey scientific findings [22, 27, 28]. The use of citations as a proxy for impact – impact, not quality – is grounded on the notion that science is built from blocks, i.e. new knowledge is constructed from previously acquired knowledge, and that peers give credit to that knowledge and recognize valuable research through citations [12, 23, 27]. Finally, according to Katz and Martin [29], even though co-authorship can only be seen as a partial indicator of collaboration, it yields many advantages, as it is invariant and easily verifiable and the fact that it is not costly, while being a practical indicator.

Table 2.2: Most used bibliometric indicators, as well as their description.

Metric	Description
Publication Counts	Number of publications produced by a researcher, research group, institution, country or journal
Citation Counts	Number of citations obtained in a publication/or total number of publications by a researcher, research group, institution, country or journal
Journal Impact Factor	Ratio between the number of citations in a given year to the publications in a journal in the previous 2 years over the number of substantive articles published in the same 2 years [30, 31]
H-index	A scholar has an index h when he has published h papers, each of which has been cited at least h times.
pptop10%	Proportion of publications of an unit of analysis that, compared with other publications in the same field and in the same year, belong to the top 10% of the most frequently cited [28]

Having in mind the main objectives of a bibliometric analysis, a set of indicators should be chosen to perform the evaluation of scientific production. Yet, one should not forget that these measures or indicators should have a meaning in the context of each study, otherwise they are futile [28, 32]. Besides that, the chosen indicators should be looked upon from a benchmarking or evolutionary perspective

[12,28], since they should provide a context to allow the evaluation of scientific production. As mentioned in the previous paragraph and presented in table 2.2, both publication and citation counts can be used as a means to evaluate scientific production.

Publication Counts

Research output can be measured through publication counts, i.e. publication counts can be a measure of productivity [33]. This indicator can result from a full counting method, in which one publication counts as one ”article unit”, or from a fractional counting method, in which ”each author is attributed a fraction of the article corresponding to their share of all authorships” [27]. However, both these methods assume that each scholar presented as an author of an article contributed in a similar fashion for said publication. In spite of this limitation for micro-level evaluation, both counting methods are ”highly correlated at the highest levels of aggregation” [27].

Productivity may be influenced by several factors, from the individual characteristics of each researcher, to the characteristics of institutions and departments, funding and even collaboration [7, 34–36]. Lee and Bozeman defend that there is a positive relationship between the acquiring of funding and scientific fields with scientific productivity of individual researchers [7]. The fact that researchers are able to capture funding also influences productivity positively is also defended by Jacob and Lefgren [36]. Furthermore, the departmental and organizational context also influences the productivity of researchers [34, 35]. Finally, collaboration may also be a factor that strongly influences productivity. The number of authors participating in publications seem to influence the productivity of individual researchers [7, 37]. Besides that, heterogeneity in the affiliations of authors, i.e. multi-institutional collaboration, also has a positive influence in productivity [7, 9, 37, 38]. International collaboration is also believed to have a positive impact in research productivity [9, 39].

Citation Counts

Besides research output, one ought to measure research impact. It is important to highlight that a higher impact does not necessarily mean that the work is of higher quality, as it may have received more attention due to other factors [27]. Having that in mind, citation counts can be used to assess attention and consequently impact [12, 28, 40]. For citation counts to be a proper metric to evaluate impact, one should understand the reason that leads scholars to cite previous work. According to Tahamtan and Bornmann, ”citing motivation is a multi-dimensional phenomenon, and scholars cite the literature for a variety of scientific and non-scientific reasons” and as such conventional citation analysis may not yield the most reliable results in terms of impact of research [40]. Previously, Harwood identified eleven reasons for citing, which are presented in table 2.3 [2].

Apart from the motives presented in table 2.3, authors may also cite a lot of publications in order to be presented as more professional [41]. Furthermore, as Wang *et al.* [41] mention, most research on reasons for citing have analysed the citing authors' opinions, ignoring subconscious reasons for citing

Table 2.3: Eleven reasons why researchers cite previous work as identified by Harwood [2].

Reasons to cite	Description
Signposting	Indicate further reading
Supporting	Justify their choices of hypothesis, methods or claims
Credit	Acknowledge previous work or ideas
Position	Identify and attribute different points of view
Engaging	Critique previous findingsor hypothesis
Building	Build from previous obtained knowledge
Tying	Align to a certain methodology, school of thought or disciplinary traditions
Advertising	Alert to previous work or data
Future	Provide suggestions of future work
Competence	Highlight aptitude in the area of study
Topical	Show that the issue they are researching is of current interest

which may be relevant to perform a citation analysis. Finally, regarding the use of citations as a measure of impact, one should not forget that authors may not cite all the publications they read that had an impact on their citing article [33].

Besides the reasons researchers may cite documents, it is important to understand what are the reasons that may contribute to a document acquiring more or less citations. First of all, differences across research fields must be accounted for, there are fields that gather more interest and consequently acquire more citations and fields that are smaller and produce less research and therefore there are less opportunities for citations [42]. However, in all fields, the capabilities researchers and institutions have to communicate their findings are of extreme importance to the acquisition of citations [12]. Besides that, collaboration, specially heterogeneous collaboration, i.e. the collaboration between different institutions, and international collaboration have a positive influence on impact [43, 44].

In producing citation analysis studies, one should always normalize citation counts, in order to account for field and disciplinary differences [31, 45]. In this normalization process, an average Field Citation Score (FCS) would be computed including all the publications produced in a given field in a given period of analysis and the ratio between citations per publication of the unit of analysis and the FCS would be calculated, thus obtaining a normalized value of the citations of a determined field [23]. Nevertheless, citation counts do not provide an ideal approach to measure research impact due to volatility and inconsistency consequently, as it happens in publication counts, it is also more adequate to characterize the research impact at higher levels of aggregation [27, 28].

As already mentioned, these indicators should provide context and as such should be analysed from a benchmarking or evolutionary perspective. A trend analysis could be a method to determine how the production and impact of a certain researcher, research group, university or even country is evolving.

Moed *et al.*. admit that three conclusions may be drawn from a trend analysis based on the number of publications and citations per publication: if the publication counts from one researcher, group, university or country is increasing, one can assume that their scientific production is increasing; if the number of citations in a certain citation window is increasing, one can deduce that the impact of that researcher, group, university or country is increasing; and if the number of citations is increasing, but the number of citations-per-publication is decreasing, it is assumed that the researcher, group, university or country have reached a saturation level [12].

Journal Impact Factor

The Journal Impact Factor (JIF) was created to compare journals not having in consideration their publication counts [24, 27, 30, 31]. Nowadays, this indicator is commonly used to assess the quality of the research produced by a single researcher, a research group or an university [46] However, this indicator has several limitations [24]. Firstly, there is an asymmetry between the numerator and the denominator. In the denominator, only citable publications are considered, yet, the citations considered in the numerator can be to both citable and non-citable publications [24, 27, 31]. Secondly, this indicator is portrayed as an average, even though it is known that citation distribution is highly skewed [24, 27, 46].Besides that, this indicator can have a too short citation window for some fields of research, in which citations only peak 3 or more years after publishing [24]. Another concern to have in mind when debating the use of this indicator is the fact that it does not allow for comparisons between journals of different disciplines, as different disciplines have different citation practises, as already mentioned [27]. However, Garfield, who was one of the developers of this indicator, defends it against these claims by saying that the period for which the impact factor is computed could easily be changed, in addition he argues that "all citation studies should be normalized to take into account variables such as field, or discipline, and citation practices" [30]. Finally, he also mentions that "it is one thing to use impact factors to compare journals and quite another to use them to compare authors" [30] and Sugimoto and Larivière conclude that this metric should only be used to assess the performance of journals [27].

H-index

To put an end to this quest of measuring the relevance of one's research output, Hirsch developed the h-index. The author characterizes this index as follows "A scientist has index h if h of his or her N_p papers have at least h citations each and the other $(N_p - h)$ papers have $\leq h$ citations each.", in which N_p is the number of publications, and he classifies it as being a "simple and useful way to characterize the scientific output of a researcher" [47]. Despite being widely used in this domain, many argue that this simplicity may be the big flaw of this index and many changes have been proposed for this indicator [24, 48]. In 2012, Waltman and Van Eck argue that "the way in which the h-index aggregates publication and citation statistics into a single number leads to inconsistent results" [48]. Besides that, Sugimoto and Larivière ackowledge that "this indicator is also prone to scientific and sociological distortions given

that it favors quantity in terms of publications" [27].

pptop10%

This indicator allows to avoid the bias of the skewness in the number of citations. It is given by the percentage of documents that are in the 10% most cited documents relative to the documents published in that field in that year. According to Van Raan, this indicator may be the most significant to measure the impact of scientific research. As it is distribution-based, instead of based on an average, it is not sensitive to outliers [28].

2.1.2 Science Mapping

As previously mentioned, bibliometrics can also be used to perform science mapping, which may be extremely useful to inform decisions in research policy. According to Moral-Muñoz *et al.*, this method is "dedicated to showing the structural and dynamic aspects of a research field, and how it evolves through time" [28]. In order to operate such mapping, bibliometric networks are built and the basic principles of such networks are as follows: these networks are constituted by nodes, actors that can depict different units of analysis, and edges, that connect those actors and depend on the kind of network generated [28,49]; edges are usually weighted and this weight depends both on the data and on the kind of network generated [49]; actors are placed according to relatedness, as such actors that are closer together are more related than actors further apart [23,50].

The focus of science mapping, like it was communicated until now, is the development of networks. Therefore science mapping is characterized by a high degree of interdisciplinary [51], with the participation of domains such as graph theory, SNA or even statistics [28] and the very important participation of computer science [51]. SNA is going to be presented in more detail in subsection 2.2.3 and the use of SNA to analyse bibliometric networks in subsubsection 2.2.3.A. These bibliographic networks enable the recognition of fields of study as well as their visualization [51,52].

These networks may have different units of analysis, such as publications, individual researchers, organizations, countries, journals and words [53]. Besides that, each of these units of analysis can be "used to compile a specific kind of structure" [23]. One can develop a citation network, a co-citation network, a bibliographic coupling network, a co-authorship network and a co-word network.

- *Citation network* - Connects documents taking into account direct citations [28,53];
- *Co-citation network* - Connects documents based on their joint appearance in the reference list of another publication [28,50,53];
- *Bibliographic coupling network* - Connects documents based on the references they share [28,50,53];

- *Co-authorship network* - Connects authors when they share authorship of a paper [53];
- *Co-word network* - Connects words based on their co-occurrence in titles, abstracts or keywords of articles [28, 50, 53].

Each kind of network is useful and adequate to measure and evaluate different features of scientific production, for example the co-authorship network may translate the social structure of a field [23], while the co-word and bibliographic network of a field may provide the cognitive structure of a field. The choice of analysis to make should be based on the research questions posed, as well as the unit of analysis one means to study [53]. The literature review for the networks to be used in this work will be presented later on.

2.1.3 Databases

Considering all the data that is necessary to perform a bibliometric analysis, as mentioned in the previous sections authors' names are needed, as well as titles, abstracts, institutions' addresses, references, the choice of Database (DB) from which to retrieve that information is of extreme relevance. To understand the bibliographic databases available today, one must first get acquainted with the Citation Index.

2.1.3.A Citation Index

Even though one can trace the measurement of science or literary production to - at least - the third century before Christ [54], the tool that allowed "the systematic study of citations", the SCI, was only created in 1964 [10, 31]. According to Garfield, the SCI "is the first really serious attempt at universal bibliographical control of science literature since the turn of the century" [19]. It is then important to understand what is a Citation Index and what are its uses, as well as to acknowledge what are the citation indexes or DBs available to perform bibliometric studies today.

According to Garfield's definition of 1964, "a citation index is an ordered list of cited articles each of which is accompanied by a list of citing articles" [19]. However, nowadays, a citation index is so much more than that. A citation index contains this citing-cited references, as well as information on the authors, affiliations, sources, funding, title of the documents, which allow for the development of bibliometric indicators at higher levels of analysis. [27, 31]. Garfield's citation index, SCI, was created not with the purpose of evaluating science, but with the purpose of facilitating retrieval of relevant articles [19, 27]. However, Garfield attributes the success of this product to "its use as an instrument for measuring scientific productivity, made possible by the advent of its by-product, the SCI Journal Citation Reports (JCR) and its Impact Factor rankings." [31]. Nowadays, the two bibliographic databases that are most commonly used to measure scientific production are the WoS and SCOPUS [28].

Web of Science

WoS is the modern successor of SCI, already mentioned in the previous paragraphs [27]. This web-based tool is "a multidisciplinary and selective DB that is composed of a variety of specialized indexes, grouped according to the type of indexed content or by theme" [33]. As already mentioned, the first use that was thought for SCI was the retrieval of scientific publications and as such since very early this index included the affiliations of the authors, so that one could search for publications of an institution or country. Because of this, WoS allowed for the analysis of collaboration networks and as such this DB is particularly appropriate for bibliometric analysis on the aggregate level [27]. Finally, it is important to understand that the coverage of this DB is specially focused on the natural and life sciences with its weakest coverage being in the social sciences and humanities [27, 33].

Some of WoS most striking features have to do with the disambiguation of institutions' names, due to its long history of standardizing data. The institutions' names variants as well as parent/child relationships between institutions are accounted for and can be found under a favored institutional name. Through the Organization Enhanced (OE) searching field, one can find all the articles of a certain institution [33]. When institutions merge or split over time, their articles even the ones anterior to the merger can be found under the new name or new unit's name [33]. Therefore, WoS may be more adequate for aggregate level bibliometric analysis.

SCOPUS

For many years, WoS was the only citation index widely available. However in 2004, that changed when Scopus was released by Elsevier [27, 33]. Just as WoS, Scopus is a citation index with bibliographic information and metadata on millions of articles and has a bigger focus on arts and humanities than the first DB presented [27], even though it is still biased towards natural and life sciences [33]. In comparison with WoS, Scopus is stronger in the disambiguation of individual authors, and as such may be more suitable for individual-level analysis [27].

2.1.3.B PubMed

The main goal of this thesis is to execute a bibliometric evaluation of the research performed by the University of Lisbon in the health domain. Having that in mind, a more specific DB than WoS or Scopus was needed to retrieve the data, and thus the use of PubMed was deemed as adequate.

PubMed is a literature DB that encompasses mainly research from medicine, life and biomedical sciences and it was introduced in 1996. This DB allows for the search of publications based on keywords and several other criteria [28, 55]. However, PubMed does not include cited references, therefore it is not a citation index [28].

The data that is present in this DB is mostly data that is covered by MEDLINE, the bibliographic DB for the life sciences with a special focus on biomedicine of the National Library of Medicine (NLM). Despite being the main datasource for PubMed, MEDLINE is not the only one, and as such PubMed

contains publications outside of the scope of this DB, primarily from chemistry and general science's journals [56].

2.1.4 Limitations and challenges of bibliometric methods

Despite being a seemingly meaningful and harmless method to perform a quantitative evaluation of scientific production, the use of bibliometrics may contribute to some questionable publication and citation practises [24, 33]. As noted by Haustein and Larivière "the more bibliometric indicators are used to evaluate research outputs and as a basis for funding and hiring decisions, the more they foster unethical behavior" and "the higher the pressure, the more academics are tempted to take shortcuts to inflate their publication and citation records" [24]. In their article, the authors present several malpractices used by researchers, institutions or journals to improve their evaluations.

Firstly, authors and institutions may be swayed away from the most adequate journals to publish their research in, because of the increasing importance of the JIF for a researcher's evaluation. This may be problematic for two reasons: the target audience may not be fully reached and at the same time important themes may not be properly investigated due to a shift from national focused research to international focused research to increase the number of publications in English speaking journals, that habitually have higher impact factors [24]. Questionable practises may also be applied to increase publication and citation counts, such as salami publishing, honorary authorship and self-citation. In the first mentioned practise, researchers may search for the smallest publishable unit in order to increase publication counts. This is unethical, because "it distorts scientific progress and wastes the time and resources of the scientific community" [24]. Honorary authorship is characterized by the attribution of authorship to highly-cited researchers, who have not participated in the investigation, so that citation counts can increase. At the same time, it can also be used to increase the publication counts of the added researchers [24]. Finally, through self-citation, authors cite their own publications, so that they can artificially increase their citation counts. At a large scale, i.e.in institutions or countries, this practise does not distort in a relevant way a bibliometric evaluation [24].

It is also important to mention that the creation of many bibliometric indicators is often unsystematic, as well as empirically based [48]. According to Waltman and Eck, "researchers take an indicator, identify a property of the indicator that they argue is undesirable, and then propose a new indicator that does not have this undesirable property" and this may lead to an inconsistent choice of bibliometric indicators [48]. This should also be considered when developing a bibliometric study. One should focus on indicators that convey production and impact in a transparent way, so that the evaluation procedure can be simplified [48].

Finally, in spite of its, already mentioned, reproducible and inexpensive approach, it may be costly in terms of time and energy. Van Raan reminds that "verification is crucial in order to remove errors

and to detect incompleteness of addresses of universities, departments, and groups, and to ensure correct assignment of publications to research groups and completeness of publication sets" and this pre-processing of data may be labour intensive [28]. The preparation of standard methods to perform a bibliometric analysis is of extreme relevance to ensure the validity of the obtained results [28].

However, with the ongoing expansion and specialization of research areas, finding experts that can characterize all the research produced becomes increasingly difficult. As such, bibliometrics provides a transparent and abragent means to perform this characterization, aiding in the development of science [42].

2.2 Collaboration

One of the objectives of this thesis is to understand how the different colleges of the UL collaborate in the production of science in the health domain. Having that in mind, it is first necessary to assert what collaboration is. As many scholars have mentioned defining collaboration is no easy task and the fuzzy boundaries may be difficult to ascertain [29, 57]. According to Katz and Martin [29], "a 'research collaboration' could be defined as the working together of researchers to achieve the common goal of producing new scientific knowledge". Besides that definition, research collaboration between institutions can be defined as "a mutually beneficial and well-defined relationship entered into by two or more organisations to achieve common goals" [58]. However, as Katz and Martin report in their publication [29], the borders of research collaboration are hard to define, since scientists partake in scientific investigation in varying degrees and consequently, so do organizations.

In spite of this difficulty in defining collaboration, it is agreed that the dimension and relevance of research collaboration has been increasing [27,28,57,59,60] and as such the study of this phenomenon grows in value. This growth is motivated by several factors that favor both researchers and institutions and sometimes even countries [29, 59–62]. Some of the main factors that benefit collaborators and, thus, drive collaboration are: the interest in acquiring new knowledge and information in a timely manner [29,59,63], the increasing need for specialization in many fields of science [27,29,59,60,63], interest in accessing expensive equipment and difficult access data [27,29,59], the easier access to funding attributed - or even demanded - by funding agencies [27,29,61,62], the broadening of the coverage of the area of expertise and consequently a stimulation of innovation [59, 61, 63], higher chances of success due to the increase of accuracy, as well as quality, that is promoted by having more minds thinking about the question of interest [59, 61, 62], the boost in interdisciplinary science that may lead to the materialization of new fields of study [29, 59], an economic development that may arise from collaborations between academia and industries [59] and, through international collaboration, an improvement of the relationships between countries [29, 59]. The benefits of collaboration are not to be taken for granted,

though. For the collaboration to be of added value to all the parts involved, it is important for it to be of success, and, as such, the reasons that promote a thriving collaboration should be accounted for. According to Bozeman *et al.* [64], ”collaborations can collapse for social reasons as well including, for example, exhausted resources, choices to redirect energies to a study viewed as more promising, or incompatibility and disagreements among collaborators”. Scholars identify five factors that may influence the relationship between collaborators and, thus, the prosperity of the collaboration: trust, stress, documentary practise, conflict and perceived success [65]. This success may be measured with objective outcomes, such as the number of publications, subjective outcomes, for example the satisfaction with the process, and the knowledge or schooling acquired from the process [66]. Sargent and Waters identify three contextual factors that may influence these outcomes. First of all, they pinpoint the fact that collaborations are not constituted solely of scholars and there may be administrative or technical staff whose support also influences the progress of the operations. Besides that, these researchers mention the importance of funding as the material resource that may affect a collaboration the most. Finally, they recognize that the climate in which this process occurs may determine its outcomes, i.e. the cultures of the participating organizations may be different and thus stimulate different strategies to deal with the research collaboration [66].

2.2.1 Co-authorship analysis

In order to understand the extension of collaboration, its benefits, as well as productivity, it is important to measure it. However, as Katz and Martin mention [29] ”only some of the more tangible aspects of a collaborative piece of work can be quantified while others most certainly cannot”. Having this in mind, one of the most used methods to measure research collaboration is co-authorship analysis [29, 57, 59, 64, 67, 68]. According to Subramanyam [67], using co-authorship to measure collaboration may be advantageous, because this method is ”invariant; easily and inexpensively ascertainable; quantifiable and non-reactive”. One cannot forget that inter-institutional collaboration does not happen only between different researchers, it may be the case that a researcher has two affiliations, as they may develop research in two different institutions, one can admit that this is, even so, inter-institutional collaboration [67]. Nevertheless, as Buknova reminds [57] ”not every research collaboration will necessarily lead to a publication and not all co-authorshiped papers are results of a collaborative research process”. These statement can be confirmed with the examples provided by Katz and Martin, for both collaborations between researchers and inter-institutional collaborations, to which co-authorship would not provide a satisfactory means to evaluate collaboration [29]. Additionally, there may exist some malpractices associated with the use of co-authorship to perform this kind of evaluation, such as ”honorary authorship” and ”ghost authorship” [24], that were already mentioned . Finally, one must not forget that ”research collaboration is a multi-dimensional process, of which co-authorship is only one potential dimension”

[68]. However, as Fonseca *et al* [63]. mentioned "The co-authorship of a technical document is an official statement of the involvement of two or more authors or organizations", and as such "co-authorship analysis is still widely used to understand and assess scientific collaboration patterns".

2.2.2 Brief Introduction to Graph Theory

In this subsection, basic, yet relevant, concepts about graph theory are going to be presented. The assessment of collaboration can be performed using SNA and several key concepts of network science have their roots in graph theory" [69], thus the focus of this section.

A graph is constituted by vertexes and edges. The mathematical formulation that represents a graph, *G*, with *V*, number of vertexes, and *E*, number of edges is *G = (V, E)*. Two vertexes that are connected by an edge, i.e. that are adjacent, are neighbours.

A graph has diverse characteristics that one should pay attention to. The degree of a vertex, *deg(v)*, is the number of edges incident on that vertex, i.e. simplified it is the number of neighbours of a vertex.

A walk is an alternating sequence of connected vertexes and a path, which is more used in SNA is a walk in which there is no repetition of vertexes. The length of a path is the number of edges in that path. The shortest path between two vertexes is the path between those edges that has the shortest length.

The distance between two vertexes is the length of the shortest path that connects both those vertexes. The diameter of a graph is the longest distance between two vertexes [70].

2.2.3 Social Network Analysis

As stated in subsection 2.1.2, co-authorship analysis can be used to assess collaboration. This web of cooperating authors can be used to build a collaboration network [71] and in the examination of this network, SNA may come in handy.

Social networks can be represented using graphs, in which actors or individuals are represented by nodes, or vertices as denominated in graph theory, and relationships are represented by links between nodes, or edges as in graph theory [69, 72]. SNA is "a theoretical perspective and a set of techniques used to understand and quantitatively measure these relationships" [63]. The use of SNA to analyse human connections may be useful and advantageous due to the fact that its main focus are relationships [62, 63, 73] and there is the belief that "structural relations are often more important for understanding observed behaviors than are such characteristics such as race, gender, age, socioeconomic status, and political ideology" [74]. Despite being highly based on graph theory, due to these real-world relationships, its metrics cannot have an uniquely theoretical or mathematical interpretation, there should be an empirical foundation for each of the statistics obtained, this is, there should be a real-world context and applicability coming out of these studies [32].

2.2.3.A Social Network Analysis of Co-authorship and Keywords co-occurrence

SNA has been widely used to study research collaboration and the structure of cognitive fields. Researchers have been applying the methods provided by SNA to bibliographic databases in order to perform citations and co-authorship analysis [62, 73] or the analysis of keywords' co-occurrence networks [22]. In the case of SNA applied to co-authorship analysis, each node is an author, organization or country and two nodes are connected by an edge if there exists a co-authored paper between them [63, 71]. When SNA is applied to keywords' co-occurrence networks, each node is a keyword and an edge between nodes exists if two keywords were associated to the same article [22]. As previously mentioned SNA enables the quantitative measuring of relationships between actors and, as such, the most used metrics in studies of co-authorship and in the study of keywords' co-occurrence networks are presented in table 2.4 (network-level metrics) and 2.5 (individual-level metrics).

A very simple graph is presented in figure 2.2, as well as the representation of nodes, edges and clusters. In the case of this graph, its size would be of 6 and the number of edges that constitute it would also be 6. Its diameter would be of 4 as it is the longest shortest path in the graph, i.e., the minimum number of steps that would be necessary to travel to reach node 6 from 2. Its density is of 0.2. Its average path length is of 1.93. The average degree of the network is 1. In this network, the likelihood of choosing a node that has a degree of 1, 2 or 3 is equal, i.e. there is a 33.3% chances that a node will have a degree of 1, 2 or 3. It has zero isolates and three clusters.

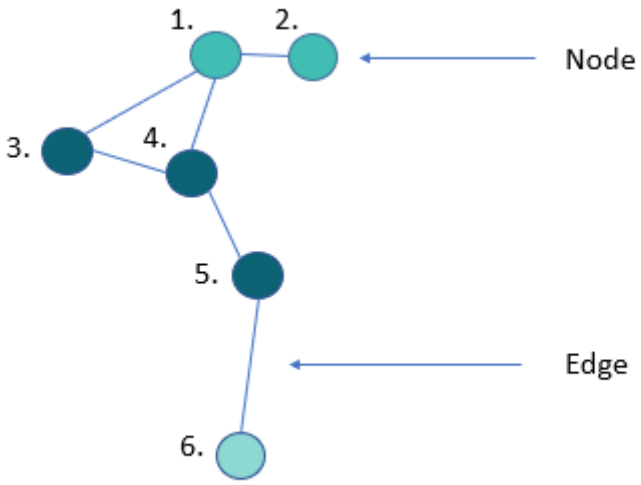

Figure 2.2: Example of a simple graph. The nodes are represented by the circles and the edges by the lines. The different colours represent clusters and, as such, the nodes 1 and 2 belong to the same cluster, the nodes 3, 4 and 5 belong to the same cluster and node 6 belongs to its own cluster. This graph has a size of 6 and has 6 links.

Table 2.4: Network level metrics used in several bibliometric studies, as well as the explanation of each one and the articles in which the metrics were used

Metrics' name	**Definition**	**Articles**
Number of nodes (size)	Number of vertexes in the graph that represents the network [69]	Aboelela et al. (2007) [72] Fonseca et al. (2016) [63] Fonseca et al. (2017) [75] Cavadas (2020) [22]
Number of links	Total number edges of the graph that represents the network [69]	Fonseca et al. (2016) [63] Fonseca et al. (2017) [75]
Diameter/Largest distance	Maximum shortest path in the network [69]	Newman (2004) [71] Cavadas (2020) [22]
Density	The number of connections the network compared to the total possible number of connections, measuring the degree of interconnectedness [22]	Aboelela et al. (2007) [72] Godely and Baron (2011) [73] Fonseca et al. (2016) [63] Fonseca et al. (2017) [75]
Average path length	Average distance between all pairs of nodes in the network [69]	Newman (2004) [71] Fonseca et al. (2017) [75] Cavadas (2020) [22]
Average degree	Average number of connections that the nodes of a given network have [75]	Fonseca et al. (2017) [75]
Degree distribution	Probability that a randomly selected node in the network has degree k [69]	Cavadas (2020) [22]
Degree Centralization	Normalised degree of the complete network [22]	Aboelela et al. (2007) [72] Fonseca et al. (2016) [63] Cavadas (2020) [22]
Isolates	Individuals that are not connected to any other in a network [73]	Haines (2011) [62] Godley and Baron (2011) [73]
Number of clusters/Clique counts	Average number of maximally connected subgroups [72]	Aboelela et al. (2007) [72] Fonseca et al. (2016) [63] Fonseca et at. (2017) [75]

Table 2.5: Individual level metrics used in several bibliometric studies, as well as the explanation of each one and the articles in which the metrics were used

Metrics' name	Definition	Articles
Degree Centrality	Number of direct connections of a node [72]	Aboelela et al. (2007) [72] Godley and Baron (2011) [73] Fonseca et al. (2016) [63] Fonseca et al. (2017) [75] Cavadas (2020) [22]
Betweenness Centrality	Degree to which a few nodes control the relationships of other noes in the network [22]	Aboelela et al. (2007) [72] Haines et al. (2011) [62] Fonseca et al. (2016) [63] Fonseca et al. (2017) [75] Cavadas (2020) [22]
Closeness Centrality	Extent of the interconnectivity of a node with all other nodes in the network [22]	Fonseca et al. (2016) [63] Cavadas (2020) [22]

2.3 Innovation

For a part of the last two centuries, innovation seemed to be promoted by a joint venture between government and firms and universities actuated as a secondary partner in this collaboration. However, the paradigm has largely changed and academic institutions play, nowadays, a very important role in the innovation process, through the triple-helix model [76, 77]. As such it is now relevant to find a means to measure and understand the evolution of innovation in a research institution, such as an institution.

First of all, there is relevancy in comprehending how innovation is produced, how new ideas are devised and new technologies arise. Innovation is thought of a constant recombination of ideas, whether from ideas that already belonged to a certain field or of ideas that may have been fetched from other fields [78, 79]. The first type of innovation is associated with "knowledge specialisation", and thus with "exploitative innovation", and the second one with "brokerage innovation", and consequently with "exploratory innovation" [79–81]. Each type of innovation presents benefits and shortcomings. Knowledge specialisation may promote an increase in efficiency and efficacy in the processes performed by an institution or firm. Despite this, a sole investment in exploitative innovation, which has limits, may leave the institution behind in acquiring new knowledge or promoting recombinant growth [79]. As funding or investment both in academic institutions and firms is not limited there must be an equilibrium in the choice of which strategy to follow [79, 80].

The knowledge already acquired by the institutions creates a knowledge network, in which the knowledge elements owned by the institution are connected based on their previous recombination [79–81]. The placement of knowledge elements in the knowledge network may influence the recombination opportunities that those elements may yet face [79].

Innovation is not a lonely activity and the position of different actors, who comprehend and own different knowledge elements, in the social network may influence the emergence of new recombinations

between elements [80,81]. This influence may be positive or negative and it may foster more exploitative or exploratory innovation. For example, Wang *et al.* found out that a researcher that presents a high degree centrality, i.e. has many connections, has less opportunities for exploratory innovation as they may be more influenced by external opinions [80].

To study these questions in the academia domain, a knowledge network must be devised. In order to do this, the co-word mapping already referred can possibly be used. Co-word mapping is used to delineate and understand the evolution of cognitive fields [82]. Cavadas used keyword co-occurence mapping to infer about the rise of new research topics in the area of marine research infrastructure [22].

3

Methodology

Contents

In this chapter, the methodology followed in this work is presented. This methodology is an adaptation from the methods performed in Cavadas (2020) [22] and Zupic and Čater (2015) [53]. Besides that, the metrics used are commonly used in most bibliometric studies.

The research design is split as follows: firstly the research questions and the methods used to answer each of those questions are presented, after that, it is explained how the data were retrieved, then the methods used to clean and standardize the data are described, following this subsection the methods to perform the analysis of the data are portrait, finally the tool used to visualize the results is presented.

3.1 Research Design

In order to develop a research project, it is important to lay out a research design that guides the project. The first step to take is to arrange the research questions and subsequently identify the proper methods to answer those questions.

3.1.1 Research Question and Methods Selection

Having in mind the aforementioned objectives, six research questions have been thought of. These questions are presented in table 3.1. The development of research questions is of extreme relevance to understand the methods with which to fulfill the objectives of any research project. In this thesis, bibliometric methods were used along with SNA.

3.1.2 Compilation of Bibliometric Data

The first step to perform after understanding the research questions is the selection of the DB for retrieval of the data. As Fonseca *et al.* mention, the DB should ”cover a large number of academic journals and have high representation of health-related journals”, ”provide information on the affiliations of the authors, allowing the construction of organizational networks” and ”the full name of the authors in most publications” and ”allow the exportation of data in text format compatible with bibliometric analysis software” [63]. Both Scopus and the WoS fulfill these criteria. However, for the reasons stated in the literature review, the fact that the names of the institutions are more concise and coherent on the latter and that the latter is more appropriate for aggregate level analysis, WoS was the chosen DB to retrieve all the articles from the UL. Besides that, to perform the identification of the articles in the health domain PubMed was used.

Having in mind the objectives devised for this thesis, publications of the UL for a sufficiently broad time period had to be retrieved. Since the UL and TUL merged in 2013, it was decided that it was best to consider only publications that were written in a posterior date. As such, the time period considered

Table 3.1: Research questions meant to be answered, as well as methods and softwares that can be used to answer them and the knowledge contribution of those answers

Research Question	Method	Software	Knowledge Contribution
How much science is produced in the University of Lisbon and what is its impact?	Bibliometrics (Publication and citation analysis)	RStudio	Quantification of the scientific production of the UL and its influence
How much of that research is performed in the Health domain and what is its impact?	Bibliometrics (Publication and citation analysis)	RStudio	Quantification of the scientific production in Health of the UL and its influence
How do colleges of the University of Lisbon collaborate when researching in the Health domain?	Bibliometrics (Co-authorship analysis) Social Network Analysis (SNA)	VOSviewer RStudio	Characterize the cooperation between the institutions of the UL
Is the research performed by one college relevant to the others?	Bibliometrics (Citation Analysis) SNA	VOSviewer RStudio	Understand if the research performed by one institution stimulates research in the others
How did the research fields in the Health domain evolved in the University of Lisbon and who promotes innovation?	Bibliometrics (Co-word Analysis) SNA	VOSviewer RStudio	Characterize the cognitive structure of the research performed in the UL and its evolution
Does the collaboration network influence the cognitive structure of the domain?	Bibliometrics (Co-word analysis Co-authorship analysis) SNA	VOSViewer RStudio	Understand if collaboration stimulates innovation or changes in the cognitive structure of the Health domain

for the evaluation was between 2014 and 2019. The evaluation was performed jointly for the period of 2014 and 2015, then for the period of 2016 and 2017 and finally for the period of 2018 and 2019 and the query for the retrieval of the data was prepared taking that into account. Besides that, in order to retrieve the publications of all the colleges of the UL the OE searching tool, whose functionality has already been mentioned, was used. The queries used in WoS were, then, as follows:

- ORGANIZATION-ENHANCED: (”Universidade de Lisboa”) Timespan: 2014-2015.
- ORGANIZATION-ENHANCED: (”Universidade de Lisboa”) Timespan: 2016-2017.
- ORGANIZATION-ENHANCED: (”Universidade de Lisboa”) Timespan: 2018-2019.

Besides that, as already mentioned, Pubmed was used to identify articles of the UL that belonged to the Health Domain. In this case, the query was as follows:

- ”Universidade de Lisboa”[Affiliation]

Adjusting the years for the time period of study.

3.1.2.A Attribution of publications to the health domain

Despite the use of PubMed to classify publications as belonging to the health domain, this method was not enough. As previously mentioned, PubMed is filled with publications from journals of the areas of medicine and the life sciences. However, the journals from these areas areas are not the sole areas that contribute to developments in health research. The social sciences, technology and even the arts and humanities can strengthen our understandings of the human body, health and treatments or cures.

Besides that, WoS classifies journals and books in Subject Categories and, then again the classifications may exclude publications that promote to research in health, but are not classified as medicine or life sciences. Using only this approach would belittle institutions whose primary research area is not health and would influence the results of this study. In order to overcome this issue, a method was devised to categorize publications that are not on Pubmed as being of the health domain, which is presented in the next section.

3.1.3 Data preprocessing

As already mentioned chapter 2, one of the most important steps in a bibliometric evaluation is the cleaning of the data. In this project there were three stages in this step of the process, firstly the standardization of the names of the institutions’ of the UL, then the articles that had been wrongly attributed to the UL were withdrawn and finally the health publications were selected.

As this thesis concentrates on the colleges of the UL, the name of these institutions' was one of the most important information fields and as such its standardization gathered special focus. To perform the standardization of the names of the institutions of the UL, the colleges or institutes that belong to the University were first identified, through the website of the UL [83]. That lead to the creation of 16 lists of alternative names for each of the sixteen institutions. After that, a thorough analysis of randomly chosen WoS's files was performed in order to capture the maximum number of variations of the names of each institution, as well as of the research units of each institution. Besides this, after the analysis lists were also created for Instituto de Medicina Molecular (iMM), Instituto de Engenharia de Sistemas e Computadores, Investigação e Desenvolvimento (INESC-ID) and the Museu Nacional de História Natural e da Ciência (MUHNAC) that were found to actively contribute to the research of the UL. Publications were assigned to these institutions when no other institution was indicated in that same entry (e.g., "UNIV LISBOA, INST SUPER TECN, INESC-ID, PORTUGAL." vs "UNIV LISBOA, INESC-ID, PORTUGAL."). After careful examination, the lists were used to perform the attribution of each publication to one or more institution of the UL with a standardized name.

After the standardization of the names and attribution of publications to each institution, there was the need to capture exclusively the publications that were worked upon by institutions of the UL and because of that only the publications that had the participation of one of the institutions' the UL or were clearly associated with the UL were captured. For the first time period, this resulted in 11900 publications from a total of 12659 (which corresponds to a wrong attribution of 6,0%), for the second time period, this resulted in 13253 publications from a total of 14168 (which corresponds to a wrong attribution of 6.5%), finally, for the third time period, this resulted in 14222 publications from a total of 15139 (which corresponds to a wrong attribution of 6.1%). Finally, the document types letters, reprints and corrections were removed from the dataset and the remaining document types were aggregated in articles, proceedings papers, reviews and others. In total, 38837 documents were included in this thesis.

After the name standardization, the data were converted to a bibliographic dataframe using the function *convert2df* from the package *bibliometrix* [84]. Then as, in WoS, there is occasionally an erroneous attribution of publications of a certain institution to another institution, there was the need to withdraw those publications of the data. To do this, a list of all the standardized names of institutions of the UL was created and any publication, that did not have one of these institutions in the authors' affiliation tag (C1) were excluded from the dataset.

Finally, it was necessary to select the publications that were related to health. This stage was a two step process. Firstly, the Digital Object Identifier (DOI)s of the publications retrieved from Pubmed were compared with those of the data from WoS. However, this step did not seem sufficient, as there are many journals in which researchers, in the health domain, of the UL may publish that are not medical or biomedical journals or similar and consequently are not indexed in Pubmed, some examples may

be publications having to do with management applied to the health setting or publications in sport's journals. Having that in mind, a thorough procedure was devised to select an appropriate number of keywords to select the remaining publications.

Firstly, there was the need to find an appropriate data source of high quality publications in the health domain. The World Health Organization (WHO) as the chosen organization for this. WHO publishes many documents relating to Health from technical documents to guidelines and these publications cover a wide range of topics, from vaccines, to devices, management of health systems and good practises. Having this in mind, it seemed it would be the best publisher to collect a considerable number of keywords related to a broad range of fields, thus solving the Health's keywords problem.

After the initial step for choosing the data source, there was the need to select the pertinent publications to capture the keywords. As already mentioned, the publications had to be characterized by a multidisciplinary and cover a wide range of topics. Besides that, the publications should also be recent in order to capture the constant evolution of the Health field. Having this in mind, the chosen publications were "WHO Guideline on self-care interventions for health and well-being" [85], "IN FOCUS: 2021" [86] and "WHO compendium of innovative health technologies for low-resource settings 2021. COVID-19 and other health priorities" [87].

To identify possible relevant keywords the three publications were read and health-related words or expressions that were not too specific were selected. After that, the words were searched for in the Merriam-Webster dictionary and the words that had any meaning outside of the scope of health were excluded. In the end, 36 root words, words or expressions were captured and both these and their possible variations were searched for in the title of the publication, in the abstract, in the authors and database associated keywords and finally in the title of the source.

The list was as follows: health, well being, hygiene, medic, vaccine, disability, self care, disease, illness, hospital, pharmac, psycholog, clinical, therap, morbidity, mortality, ageing, death, epidemi, pandemic, biocompatib, quality of life, infect, anatom, pharmac, physiologic, blood, dental, primary care, secondary care, tertiary care, intensive care, critical care, bone, soft tissue, surgical. This was an experimental method and, as such, required validation. To perform said validation, the sensitivity (equation 3.1), specificity (equation 3.2) and accuracy (equation 3.3) of the method were calculated [88].

$$Sensitivity = \frac{TP}{FN + TP} \tag{3.1}$$

$$Specificity = \frac{TN}{FP + TN} \tag{3.2}$$

$$Accuracy = \frac{TP + TN}{FN + FP + TP + TN} \tag{3.3}$$

In order to compute these values, a sample of one hundred articles were obtained both from the

Table 3.2: Name, version and description of the software used.

Software used		Version	Description
R and R-studio (packages)	bibliometrix [89]	3.1.4	Bibliometrix is an R-package specifically created for bibliometric analysis.It can be used with data retrieved from WoS, SCOPUS, Digital Science Dimensions, The Lens, CDSR and PubMed. Besides computing most analytical bibliometric indicators, it is also a tool to perform science mapping, allowing for the creation of citation, co-citation. bibliographic coupling, collaboration and keyword co-occurrences networks.
	sna [90]	2.6	This package was created to facilitate the installation of the various statnet packages that aid in social network analysis. Used jointly, these packages allow for the integration, visualization, exploration and analysis of networks.
	ggplot2 [91]	3.3.5	This package allows for the creation and visualization of graphics based on the grammar of graphics.
VOSviewer [92]		1.6.16	VOSviewer is a free software that can be used to create and visualize bibliometric networks.

group of articles captured by the keywords and from the remaining group, regarding the three time periods resulting in a total number of six hundred articles, and then both groups were carefully analysed. For the first group of articles, the goal was to estimate the number of articles that did not refer to the Health domain and were wrongly attributed to that group. For the second group of articles, the goal was to estimate the number of articles that referred to the Health domain and were mistaken as not pertaining to this domain. After this analysis, the values for sensitivity, specificity and accuracy of this method were of 0.85, 0.86 and 0.85, respectively. The method was deemed as sufficient.

Using this method and the identifiers obtained through PubMed, the number of health publications retrieved was of 4102 (1659 from PubMed and 2443 using the described method) for the first time period, 4662 (2495 from PubMed and 2167 using the described method) for the second time period and 5947 (2996 from PubMed and 2951 using the described method) for the third time period.

3.1.4 Analysis

After the data pre-processing, it was finally possible to obtain the results and perform the analysis. The software used to perform this analysis can be found in table 3.2. In this table the most relevant R-packages used are presented, as well as the visualization software that was chosen.

The data were very diverse and referred to both several units of aggregation (institutions from the UL and the UL as a whole) and to different subjects (health and all the subjects in which the UL produces science) and as such the computation of the results was performed in steps. The chosen metrics to

present are described in the next paragraphs.

Firstly, to evaluate the productivity of the UL, as already mentioned, publication counts were performed. This value was presented for the UL both for health and for all subjects, and for each institution of the UL considered, in the same manner, for health and for all subjects. The variation in the number of publications among the different time periods was also calculated. Besides that, as not all publications are the same and contribute the same to the production of scientific knowledge, the document type of each publication was considered, both for health and for all subjects, and both for the UL and for each institution.

Furthermore, each publication was attributed to a Research Area so that the influence of different areas in the scientific research produced by the UL could be described. Publications were attributed to Research Areas based on the attributions the WoS makes of each subject category [93]. As such, publications were split into ”Arts and Humanities”, ”Physical Sciences”, ”Social Sciences” and ”Technology”. The research area of ”Life Sciences and Biomedicine” was further divided in ”Life Sciences” and ”Medicine”. WoS attributes journals and books to at least one subject category, therefore, publications are classified into one or more subject categories, and consequently research areas, grounded on the source they were published on.

The keywords were also studied. In this study, only the keywords associated by the authors to their publications. The number of keywords and its variation and the most frequent keywords were presented both for the health domain and for all domains, and for the UL and each institution. These could provide an insight into the cognitive field of the areas that are researched in the UL and each institution and could also help study innovation.

To study the impact of the publications of the UL, citation counts were performed. The number of citations, as well as its distribution, the average number of citations per year, the number and percentage of non-cited papers and the variation of these metrics are presented for the health domain and for all subjects for both the UL and each institution.

Finally, regarding the part of descriptive bibliometrics, to evaluate collaboration, the number of authors, the median of number of authors per article and the percentage of single-authored articles are accounted for. Besides this, the number of collaborating institutions and variation are also presented and finally the number of collaborating countries and the top 10 of most collaborative countries in the UL collaboration network. These metrics are presented for both the health domain and all subjects and for the UL.

After the descriptive bibliometric results, there was the need to analyse the networks created by the collaboration between institutions and the co-occurrence of keywords, i.e. the social structure and cognitive structure of the research performed by the UL, respectively. These networks were only analysed for the health domain. From these networks, a number of metrics that could aid in the analysis of col-

laboration and innovation were computed. From the metrics proposed in the literature review, only a few deemed as the most relevant were calculated, namely, the number of nodes, the number of links, the average path length, the diameter, the density and the average degree. For the most influential nodes, the number of clusters was also studied. For the collaboration network, the degree centrality of the institutions of the UL was also computed.

When talking about collaboration, the number of nodes can provide insight related to the capture of new partnerships or loss of old ones. The number of links can aid in understanding how intra-network collaboration is increasing or decreasing. The average path length, the diameter and the density are all measures of interconnectedness inside the network. A smaller average path length and diameter indicate that the distance between nodes in the network is decreasing, i.e. gaps in collaboration are bridging, while a higher density indicates that elements are becoming more connected. The average degree indicates in how many collaboration relationships the average actor in the network participates in. Finally, clusters indicate similarity between authors, and as such two actors belonging to the same cluster in a network may indicate that they collaborate with each other more often or that they collaborate more frequently with the same institutions.

Regarding the keyword co-occurrence network, the number of nodes may indicate an expansion or retraction of the cognitive field and the acquisition or loss of knowledge elements. The number of links in the network can be used to understand to what extent exploitative innovation is being performed. The average path length, the diameter and the density can aid in the comprehension of how interconnected the knowledge elements are. The average degree indicates with how many other knowledge elements, the average knowledge element is paired with, consequently an increase of this metric, indicates that the average element is being recombined with more elements.

3.1.5 Visualization

The visualization of the networks can provide insights to both specialists and people who are not in the field, as the visualization is more intuitive than numbers. As such, a software whose primary objective is to create bibliographic networks was used, VOSviewer. VOSviewer was created by van Eck and Waltman from the Leiden University and is freely available for download [94]. In this thesis, VOSviewer was used through the R-package bibliometrix.

4

Results

Contents

The aim of this thesis is to portray the production of science by the UL, especially in the health domain. The production of scientific publications is characterized by several aspects, namely productivity, impact and quality, fields of study, collaboration and innovation. To characterize these aspects of the research of the UL, the methods depicted in the chapter 2.3 were followed and the obtained results are presented in this section. This chapter is organized as follows: firstly, the descriptive data obtained from a bibliometric analysis are presented, followed by the data obtained from the merging of bibliometric methods with SNA.

4.1 Productivity, Content, Innovation, Impact and Collaboration - Descriptive Bibliometrics

In this section the data obtained from the bibliometric analysis is presented. Firstly, the research produced by the UL as a whole is described, both in the generality of publications and for the health domain. In this subsection, the number of publications and document types, the research areas and keywords, the number of citations, the number of authors, institutions and countries, as well as the evolution of all these data are presented. Secondly, the research produced by each institution of the UL both in the generality of publications and in the health domain can be found. In this subsection, the number of publications, the keywords and the impact are presented.

4.1.1 Research of the University of Lisbon

As already mentioned, the research of the UL was studied for three consecutive time periods. This description began with productivity measures, such as publication counts, type of documents published and the subject categories that were researched in the UL. In this subsection, the results will be presented simultaneously to the generality of the research of the UL and to the publications previously attributed to the health domain, so that a comparison between the two can be drawn.

Number of Publications and Document Type

In figure 4.1, it is observed that the number of publications produced by the UL has been increasing. From the first time period to the second, there was an increase of 11.3% in the number of publications produced by the UL. From the second time period to the third, there was an increase of 6.7% in the number of publications produced by the UL. The health domain presented a more significant growth in the number of publications, from 2014/15 to 2016/17, there was an increase of 13.6% of the publications produced by the UL in this domain, followed by a greater growth from the second time period to the third one by 27.6%. It can also be noticed that publications classified as being of the health domain constitute more than one third of the total publications produced and indexed by the UL (34.9% in 2014/15, 35.6%

in the second time period and 42.6% in the third time period).

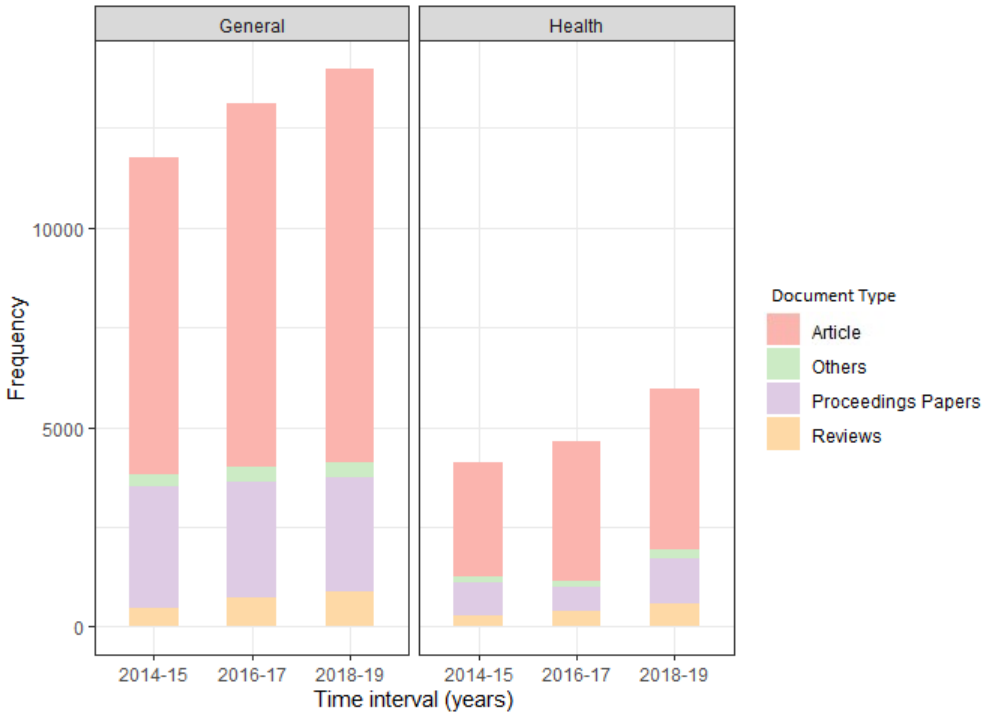

Figure 4.1: Number of documents produced by the UL for each of the studied time periods in this work, both in the generality of publications (left panel) and in the health domain (right panel), whose data are a subset of the left-side data. The proportion of publications that are classified as articles, proceedings papers and reviews can be found, as well as the proportion of publications that do not fit any of these categories (Others). It can be seen that publications in the health domain represent more than a third of all the publications of the UL

As can be seen in table 4.1, the document type that contributes the most to scientific production in the UL is the article. This document constitutes more than 67% of the publications, for the three time periods in study, and both for the generality of research produced in the UL and the health domain. Proceedings papers are the second most used document type to convey findings and in third place reviews can be found. The increase in the number of publications for the generality of research produced at the UL can be attributed to a growth in the publications of articles (14.3% from the first to the second time period and 8.3% from the second to the third time period) and of reviews (60.1% from the first to the second time period and 19.6% from the second to the third time period). However, the growth in the health domain is not so linearly explained. Reviews present a constant growth in both transitions (44.0%). In the first time period, articles present a growth of 24%, yet the greatest difference is observed for proceedings papers from the second to the third time period (87.6%).

Research Areas

In order to understand how the UL produces scientific research, one should determine the research areas in which the University publishes. In figure 4.2, the percentage of publications that are assigned

Table 4.1: Absolute number of publications and relative contribution of each document type for the total of the number of publications (in parenthesis) for the time periods in study in this work, both for the generality of publications of the UL and for the health domain. The publications of the health domain are a subset of the generality of publications of the UL

	2014/15		2016/17		2018/19	
	General	Health	General	Health	General	Health
Articles	7961 (67.6%)	2842 (69.3%)	9097 (69.4%)	3523 (75.5%)	9854 (70.5%)	4028 (67.7%)
Proceedings Papers	3066 (26.1%)	841 (20.5%)	2896 (22.2%)	595 (12.8%)	2833 (20.3%)	1116 (18.8%)
Reviews	466 (4.0%)	280 (6.8%)	746 (5.7%)	404 (8.7%)	892 (6.4%)	583 (9.8%)
Others	273 (2.3%)	139 (3.4%)	359 (2.7%)	140 (3.0%)	391 (2.8%)	217 (3.7%)
Total	11766	4102	13098	4662	13973	5947

to the different research areas are presented, both for the generality of publications of the UL and for the health domain.

In the left panel of figure 4.2, it can be observed that a considerable amount (above 30%) of the publications of the University are attributed to the research area of ”Technology”. A significant amount of the publications (above 20%) are attributed to the field of ”Physical Sciences”. However, both of these fields show a slight sign of decline, while both ”Life Sciences” and ”Social Sciences” show signs of a steady increase in the contribution to the publications of the UL. The ”Arts and Humanities” whose contribution to the publications of the UL indexed by WoS is the smallest do not show a clear pattern of growth, even though their percentage increases in both the second and third time period relatively to the first time period.

In the right panel of figure 4.2, the relative contribution of the distinct research areas to the research performed in the health domain can be found. As would be expected, a majority of the articles are classified in the field of ”Medicine” (above 35%), followed by ”Life Sciences” (above 25%). The relative contribution of the area of technology is much smaller than in the generality of publications of the UL (below 16%). The only area that shows a clear tendency of growth, even if narrow, is the area of ”Social Sciences”.

Keywords

Although the study of the Research Areas investigated by the UL already provides some information about the themes and topics preferred by the researchers of the University, the study of the keywords attributed to the publications by their authors allows the introduction of more dimensions to this picture. In table 4.2, it can be seen that in the periods of study there was a considerable increase in the number of unique keywords used by researchers in their publications in both the generality of publications produced by the UL and in the health domain of these publications. Besides that, the ratio between the number of

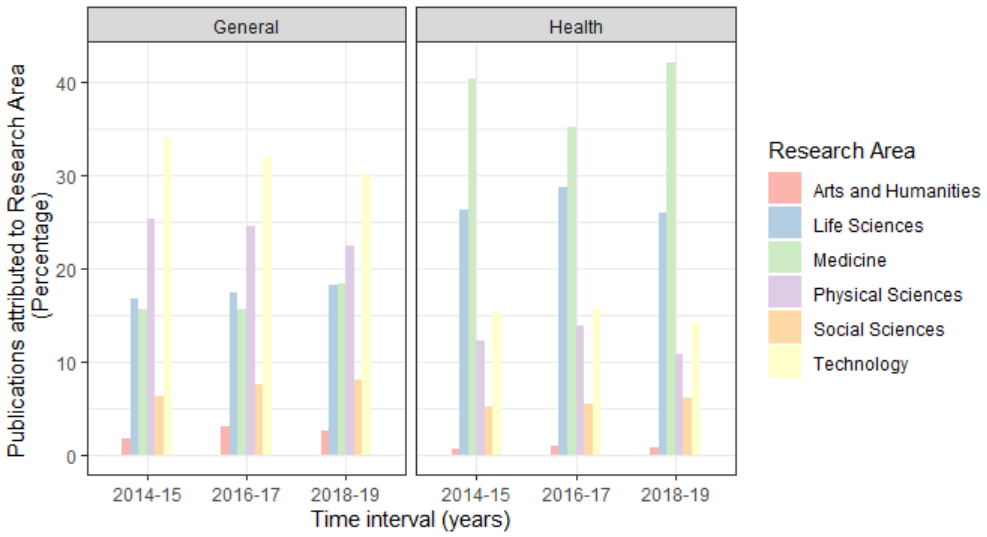

Figure 4.2: Percentage of publications produced by the UL attributed to each research area for each of the studied time periods in this work, both in the generality of publications (left panel) and in the health domain (right panel), whose data are a subset of the left-side data. Research areas were defined according to WoS's classification with the split of "Life Sciences and Biomedicine" in "Life Sciences" and "Medicine". The attribution of publications to each Research Area has to do with the classification of the journals or books in which they are published, and as such, a publication can figure in one or more Research Areas.

unique keywords and the number of documents increases for the generality of the publications of the UL. However, this ratio does not present a clear tendency when one focuses on the health domain, despite showing a more significant increase from the first time period to the second.

The keywords that researchers of the UL from all fields most frequently associated to their publications were the ones present in table 4.3. It can be observed that more than 1% of the documents produced by the UL use the keyword "Portugal". Besides that, it can be observed that there is a considerable increase in the number of articles that approach the theme of Climate Change, with the use of that keyword and "Sustainability". Keywords that are most related to the health domain appear in the ranking in the second time period, with "Parkinson's Disease", and in the last time period in study

Table 4.2: Number of unique keywords associated by the authors to the publications of the UL, as well as the variation of the absolute number of unique keywords, in percentage, and ratio of unique keywords per document, for the time periods studied in this project, both for the generality of publications of the UL and for the publications of this institution in the health domain, which are a subset of the general publications.

	General			Health		
	Number of Keywords	Variation (%)	Number of Keywords per Document	Number of Keywords	Variation (%)	Number of Keywords per Document
2014/15	26149		2.2	9392		2.3
2016/17	30255	+15.7	2.3	11950	+27.2	2.6
2018/19	33437	+10.5	2.4	14289	+19.6	2.4

”Cancer” also enters the top 10 most used keywords in the publications of the UL.

Table 4.3: 10 most frequent keywords that authors associated to the publications of the UL and the frequency of use for the generality of the publications for the three time periods studied in this project

2014/15		2016/17		2018/19	
Keyword	Frequency	Keyword	Frequency	Keyword	Frequency
Portugal	176	Portugal	226	Portugal	215
Hadron-Hadron Scattering	55	Jet	64	Climate Change	83
Climate Change	49	Climate Change	59	Sustainability	71
Optimization	42	Hadron-Hadron Scattering (Experiments)	48	Machine Learning	53
Simulation	36	Optimization	40	Hadron-Hadron Scattering (Experiments)	52
Physical Activity	30	Parkinson’s Disease	38	Optimization	52
Concrete	29	Children	34	Uncertainty	43
Sustainability	29	Sustainability	34	Cancer	42
Finite Element Method	28	Nanoparticles	30	Jet	41
Galaxies: Evolution	28	Finite Element Method	29	Parkinson’s Disease	35

In table 4.4, the most frequent keywords associated by authors to the publications of the health domain from the UL can be found. As it had already happened in the generality of the publications, ”Portugal” is the most used keyword, figuring in more than 1.5% of the publications produced by the UL.

Impact

Impact is an important characteristic of scientific production. As mentioned in subsection 2.1.1, citation counts, the proxy for impact, are relevant to understand the attention that is being given to publications. In other words, it may be used to assess how good are institutions or researchers at communicating their findings and how relevant are these findings for the scientific community. As such, it was also analysed in this study.

In figure 4.3, the distribution of the citations obtained by the UL in the time period in study is presented. A decrease in the number of citations can be observed both for the generality of the publications and for the publications of the health domain. Between 2014 and 2015, the median of citations for the generality of the publications was of 7 [interquartile range: 1;19] and for the health domain it was of 11 [3;25]. For the second time period, it was of 5 [1;15], for the general domain, and of 9 [3;20], for the health domain. Finally, for the last time period, the median of the citations of the generality of the publications was of 3 [0;9] and of 5 [1;12] for the health domain. It can be observed that most publications receive few citations, while some outliers receive many. Besides that, excluding the outliers, most

Table 4.4: 10 most frequent keywords that authors associated to the publications of the UL and the frequency of use for the publications of the health domain for the three time periods studied in this project

2014/15		2016/17		2018/19	
Keyword	Frequency	Keyword	Frequency	Keyword	Frequency
Portugal	83	Portugal	88	Portugal	95
Physical Activity	26	Parkinson's Disease	38	Cancer	41
Obesity	22	Alzheimer's Disease	28	Parkinson's Disease	35
Parkinson's Disease	22	Oxidative Stress	28	Quality of Life	53
Adolescents	21	Epidemiology	25	Obesity	32
Alzheimer's Disease	20	Quality of Life	25	Epidemiology	31
Quality of Life	20	Children	23	Alzheimer's Disease	30
Climate Change	19	Obesity	23	Climate Change	29
Epidemiology	19	Systematic Review	23	Stroke	29
Malaria	19	Amyotrophic Lateral Sclerosis	20	Aging	27

publications of the health domain receive more citations than the generality of publications.

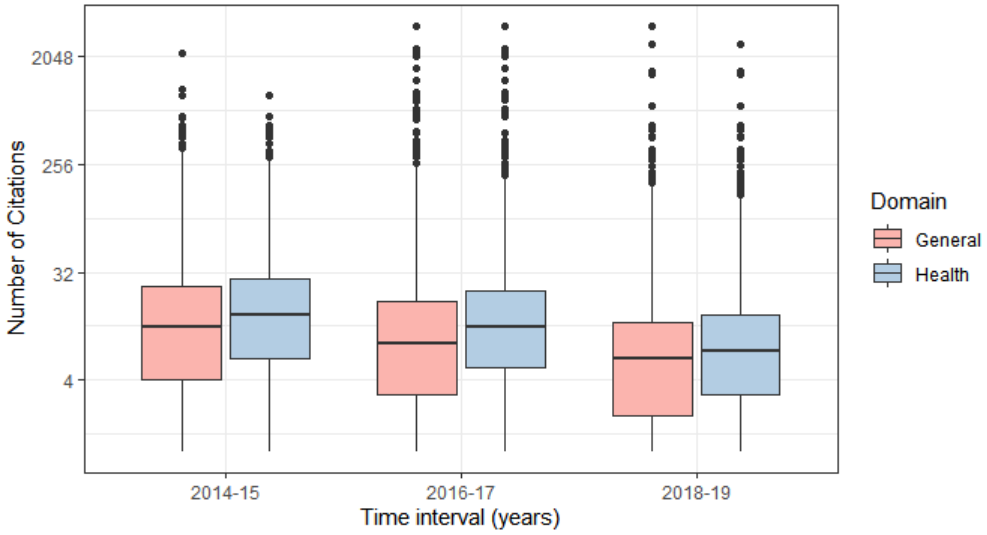

Figure 4.3: Distribution of the number of citations acquired by the publications of the UL in the three time periods studied in this project, both for the generality of publications and for the publications (in pink) of the health domain, which is a subset of the general publications data, (in blue), in a logarithmic scale of base 2.

A more accurate way to depict citations may be the average number of citations per year, which

allows for a better comparison between time periods, which is presented in figure 4.4. In this case, an increment of the average number of citations per document per year can be observed from the first time period to the second one, followed by a slight decline to the third time period, both for the generality of publications and publications of the health domain. It can also be observed that documents in the health domain receive on average more citations per document per year than the generality of publications.

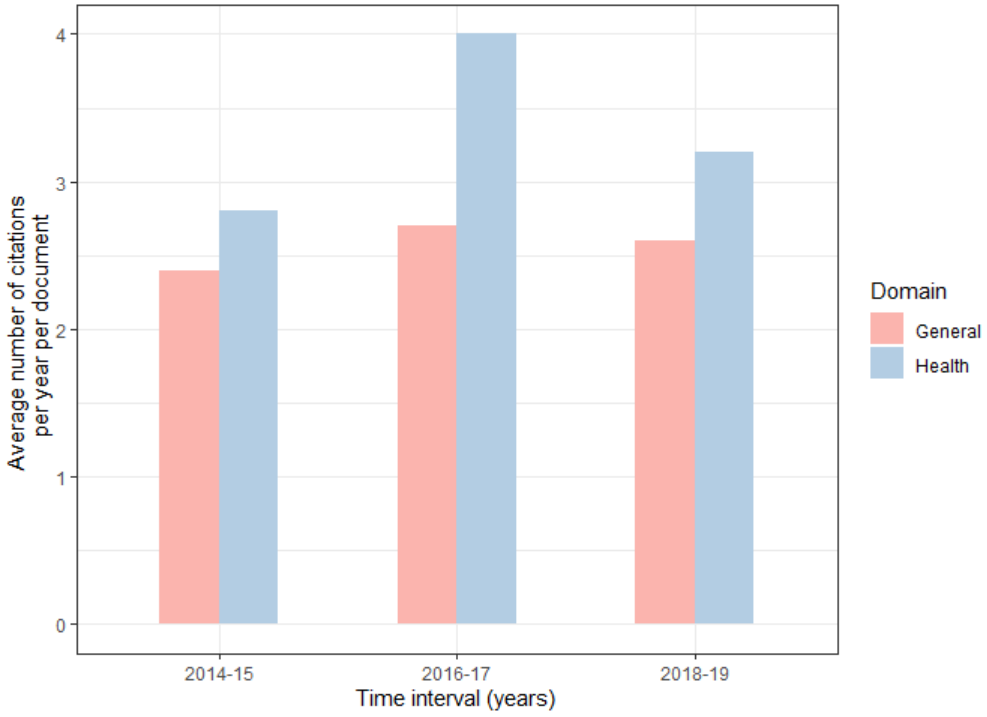

Figure 4.4: Average number of citations received by publications of the UL normalized to the number of documents of that year and to the years that have passed since publication for each time period in study, for both the generality of publications (in blue) and publications of the health domain, which are a subset of the general publications, (in pink)

When talking about citations and impact, acknowledging the percentage of documents that were not cited is necessary. In the first time period, there are 19.6% of the generality of publications that have not been cited yet and 19.4% of the publications of the health domain that are in the same situation. For the publications of the second time period, there are 20.9% of the general domain that have not been cited yet and 13.4% of the publications in the health domain that have not acquired citations so far. Finally, between 2018 and 2019, 25.9% of the generality of publications have not been cited yet and 24.4% of the publications in the health domain that have acquired zero citations so far.

Authors

To start drawing the picture that depicts the evolution of collaboration, one should start to look upon the individual researcher, as the collaboration process begins with this actor. In figure 4.5, one can study the number of authors involved in publications of the UL. It is observed that there was a significant

growth in the number of authors both from the first time period to the second (31.4% for the generality of publications and 26.6% for authors that produce research in the health domain) and from the second to the third time period (28.4% for the generality of publications and 52.2% for the health domain). Besides that, in the figure we can observe that more than 50% of the researchers that collaborate in publications produced by the UL also co-author in the health domain.

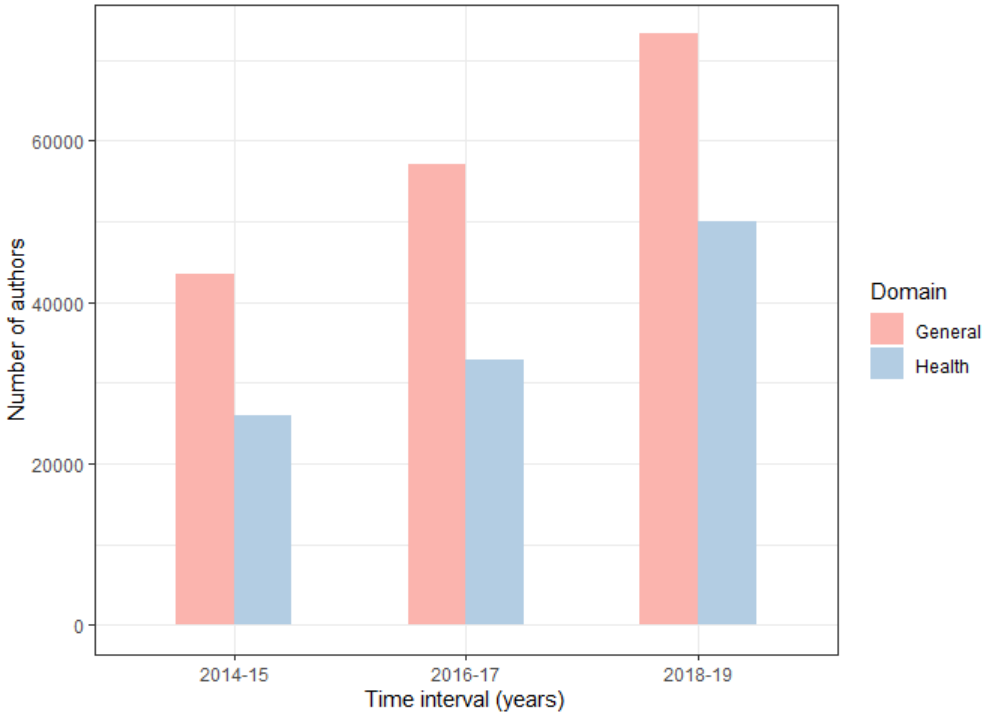

Figure 4.5: Number of authors that co-authored a publication of the UL for each time period in study, for both the generality of publications (in blue) and publications of the health domain, which are a subset of the all publications, (in pink)

Furthermore, in order to study collaboration, one should be aware of how those absolute numbers translate to the participation in publications. In table 4.5, the median number of authors per paper can be seen, as well as the percentage of single-authored papers produced in each time period. From the studied data, there is no clear trend regarding the authorship of publications. However, it can be noticed that publications in the health domain are characterized by a higher degree of co-authorship than the generality of publications.

Institution

In order to analyse collaboration between institutions, it is necessary to assess the amount of institutions that have participated in publications produced by the UL. In figure 4.6, it can be observed that the number of institutions taking part in documents published by the UL has been increasing, both for all publications and for publications in the health domain. From the first time period to the second, this growth was of 37.9%, followed by an increase of 44.5% from the second to the last studied time period,

Table 4.5: Median number of participating authors per publication and percentage of single-authored document for the three time periods in study, both for the generality of publications and for publications of the health domain.

	General		Health	
	Median	Single-authored (%)	Median	Single-authored (%)
2014/15	4	5.8	6	3.0
2016/17	4	7.3	6	3.0
2018/19	5	6.4	6	2.9

for all publications. This surge was even more substantial in the health domain, in which there was an increase of 49.6% of the institutions in the first transition between periods and of 63.8% in the second transition between periods.

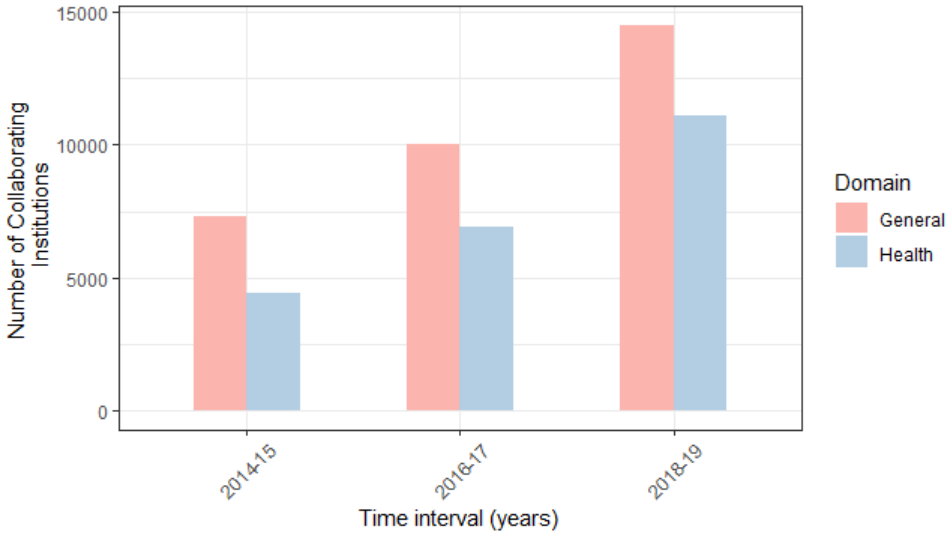

Figure 4.6: Number of unique institutions that participate in publications produced by the UL in the times period studied in this project, both for all publications (in pink) and for publications in the health domain (in blue), which are a subset of the data on the left.

Countries

Finally, collaboration is not exclusive to researchers or institutions. Scientific collaboration between countries can improve diplomatic relationships and may be a way to mitigate inequalities, as has already been mentioned in section 2.2, but above all its increase may be an indication of the growing internationalization of research. As such, to characterize the scientific production of the UL the partnerships formed with institutions from other countries should also be studied.

Intuitively, the researchers from the UL collaborate more frequently with other researchers from Portugal. Despite this, for the generality of publications, international collaboration experienced a slight growth from the first time period to the second time period, from 75 collaborating countries to 85 coun-

tries. In the third time period, there was a maintenance of the number of collaborating countries. In the health domain, the number of countries participating in health publications of the UL increased by more than 16% in the transitions between time periods. In the first time period in study, there were 55 countries collaborating in publications of the UL, in the second time period there were 64 and in the third 74 countries were participating in the health publications produced by the UL.

In table 4.6, the countries that collaborated the most with the UL in the generality of the publications are presented. Their relative contribution over the years has been increasing. From 2014/15 to 2016/17, their contribution grows from 18.5% to 20.9%. From 2016/17 to 2018/19, this value increases to 22.5%. Countries from the European Union (EU) represent more than 55% of the collaborations in the top most collaborative countries, for the general domain.

Table 4.6: Countries whose institutions collaborate the most in the generality of publications of the UL and their absolute contribution for the three time periods studied in this project.

2014/15		2016/17		2018/19	
Country	Number of Publications	Country	Number of Publications	Country	Number of Publications
France	373	UK	418	UK	559
Spain	344	Brazil	370	Brazil	484
UK	325	Spain	366	Spain	437
USA	279	France	361	USA	395
Brazil	220	USA	357	Italy	292
Germany	220	Germany	305	France	287
Italy	196	Italy	240	Germany	256
China	121	China	194	China	249
Switzerland	100	Morocco	121	Morocco	184

In the health domain, the relative contribution of the most collaborative countries does not follow a clear trend. There is a considerable increase from the first time period to the second, from 16.1% to 20.9%. However, there is a slim decrease to 20.3% in the third time period in study. The contribution of countries from the EU is more significant than in the general domain and it oscillates between 66.7% and 77.8%.

4.1.2 Research of the Colleges of the University of Lisbon

As the UL is a university constituted by many colleges, institutes and research units, its scientific production cannot be characterized without understanding the context of the research in each of these institutions. As such, in this subsection, the number of publications and its trend, as well as the most common keywords common to the publications of the institutions, and impact are presented for the instituions of the UL, both for all their publications and their publications in the health domain. **Number of**

Table 4.7: Countries whose institutions collaborate the most in the publications of the health domain of the UL and their absolute contribution for the three time periods studied in this project.

2014/15		2016/17		2018/19	
Country	Number of Publications	Country	Number of Publications	Country	Number of Publications
UK	120	UK	171	UK	246
USA	110	USA	153	USA	198
France	96	France	143	Brazil	181
Spain	92	Brazil	119	Spain	153
Germany	67	Spain	113	Italy	110
Brazil	52	Germany	105	France	107
Italy	48	Italy	60	Germany	95
Switzerland	38	Belgium	56	Netherlands	59
Netherlands	36	Netherlands	53	Australia	56

Publications

Production can be measured through the number of publications. In this case it is important to assess both the relative contribution of each institution to the research produced by the UL and the evolution of the number of documents published by each institution. It can be understood that there are three distinct groups in terms of scientific production that is indexed in the WoS. As can be observed in figure 4.7, there is a highly productive group constituted by IST, FCUL and FMUL, whose summed relative contribution constitutes approximately 80% of the publications of the university. This group is followed by a moderately productive group which is formed by iMM, ISA and FFUL. Finally, there is a group of institutions whose contribution to the scientific production of the UL is much lower.

In the highly productive group, only FMUL presented a considerable growth in the number of publications in the time periods in study. In both transitions, the number of documents produced by this institution grew by more than 20%.

In the group of moderately productive institutions, both ISA and FFUL go through a transition of substantial increase in the number of publications they produce. ISA faces an increase of 24.8%, while FFUL sees a soar in its production of 45.9%.

It is in the group of least productive institutions that the more diverse changes happen. The institutions that present a considerable rise in the number of publications in the first transition are: FBAUL by 166.7%, ISCSP by 85.7%, FLUL by 85.6%, FPUL by 60.6%, FMV by 43.0%, Instituto de Ciências Sociais (ISC) by 40.8%, FMD by 33.3%, MUHNAC by 30%, IE by 25.0% and IGOT by 21.0%. In the second transition, the institutions whose production grew the most were: FMD by 275.0%, FDUL by 150.0%, FPUL by 58.2%, FAUL by 43.3%, IGOT by 36.2%, ISC by 35.5%, ISCSP by 32.0%, FLUL by 30.6%, Instituto Dom Luiz (IDL) (by 26.9%) and FBAUL by 25.0%. However, there were two institutions whose production decreased between the second and the third time periods in study. FMV production declined by 20.2%, while the production of MUHNAC decreased by 23.3%.

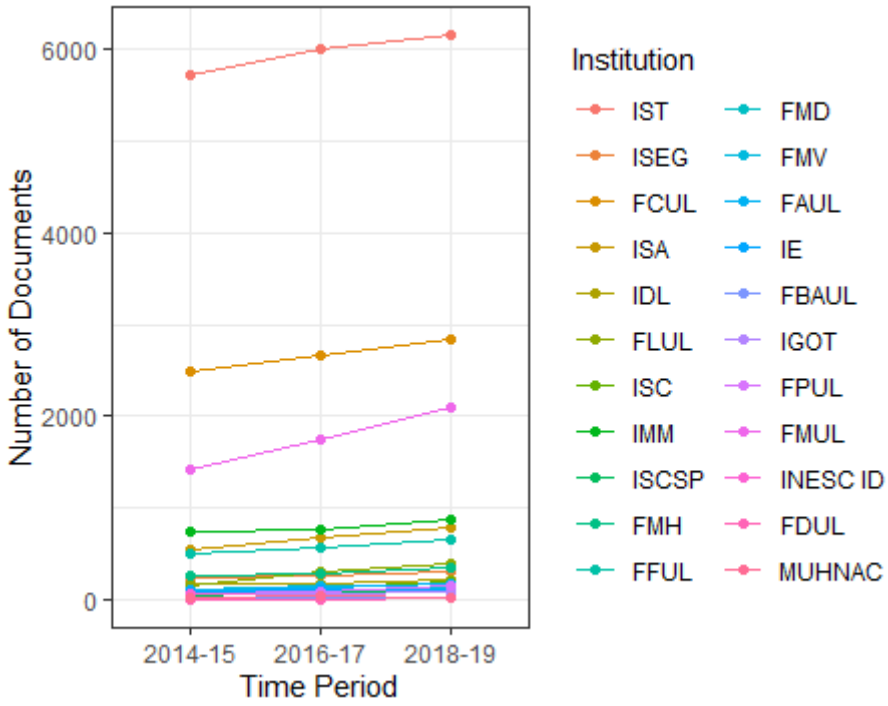

Figure 4.7: Evolution of the production of all publications of the institutions of the UL in the three periods of time considered in this project.

Regarding the health domain, the production of publications by the different institutions is much more homogeneous. In this case, the four most productive institutions contribute to more than 80% of the documents authored by the UL, IST, FMUL, FCUL and iMM.

As can be seen in figure 4.8, most institutions present an ascending trend in the number of documents produced in the health domain. In the first time period, these institutions experienced a significant increase in the number of publications: IDL by 233.3%, ISCSP by 107.1%, IGOT by 80.0%, ISA by 74.7%, IE by 60.0%, FPUL by 55.3%, ISEG by 52.2%, FMD by 50.0%, FMH by 33.5%, MUHNAC by 33.0%, FMUL by 28.8%, FFUL by 24.6%, IST by 24.6%, FMV by 23.2% and FCUL by 22.4%. In the second time period, the institutions who presented the most meaningful increase were: FAUL by 135.0%, IE by 118.8%, ISC by 85.2%, FPUL by 76.3%, FLUL by 73.2%, ISCSP by 41.4%, ISEG by 40.0%, IGOT by 38.9% and FFUL by 20.6%. Despite the general tendency to grow, there were four institutions that presented a decline in the number of publications in the first and/or the second transition. The number of documents produced by FDUL decreased in both the transitions between time periods by 50.0% in the first one and by 100.0% in the second. FBAUL, MUHNAC and FMV, presented a drop in the number of publications produced by 33.3%, 25.0% and 6.4%, respectively.

Keywords

As previously stated, the keywords associated to the publications may be useful to define the themes

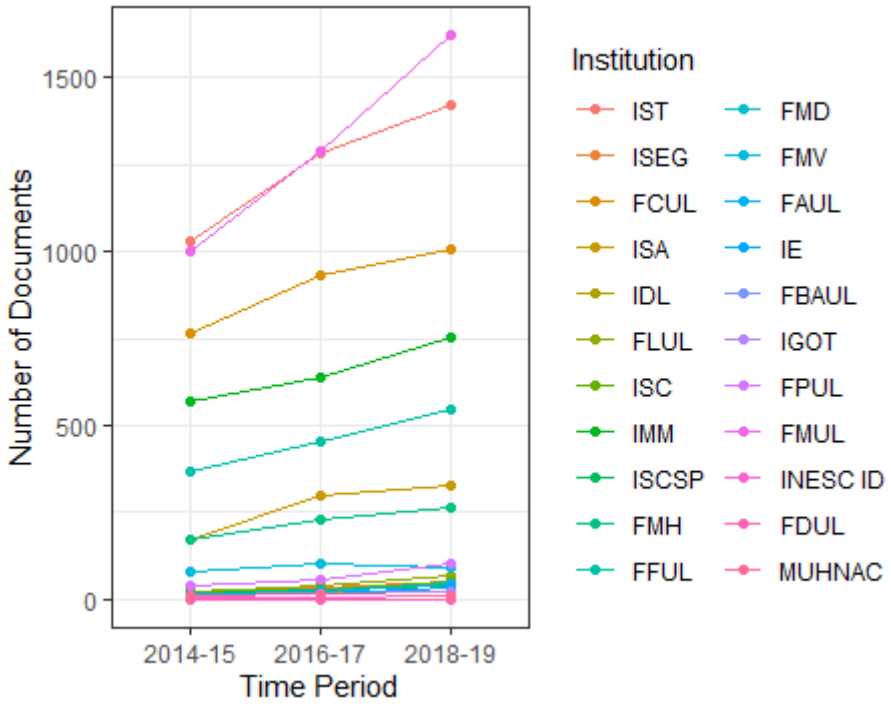

Figure 4.8: Evolution of the production of publications of the health domain of the institutions of the UL in the three periods of time considered in this project.

investigated by the research produced by the institutions, in a very broad sense. It may also be useful to identify possible opportunities for collaboration. As such, in the appendix, a table of the most frequently used keywords of each institution for each of the time periods in study can be found.

For the general domain, between 2014 and 2015, the keywords that were most used by each institution and that were used by different institutions were ”hadron-hadron scattering”, ”Parkinson’s disease”, ”obesity”, ”quality of life” and ”virtual reality”. In the second time period, the keywords that were most used by each institution and that could be found on the ranking of different institutions were ”Parkinson’s disease”, ”amyotrophic lateral sclerosis”, ”obesity”, ”rheumatoid arthritis”, ”climate change”, ”MODIS” and ”sauropoda”. Finally, in the last time period, in the generality of publications, the most used keywords that are frequently used by more than one institution were ”hadron-hadron scattering”, ”Parkinson’s disease”, ”climate change”, ”sustainability”, ”ecosystem services”, ”biodiversity”, ”Parkinson’s disease”, ”aging”, ”obesity”, ”quality of life”, ”amyotrophic lateral sclerosis”, ”cancer” and ”exercise”.

In the health domain, the most common keywords used by each institution that are frequent in each other’s rankings, in 2014 and 2015, were ”Parkinson’s disease”, ”nanoparticles”, ”Alzheimer’s disease”, ”climate change”, ”obesity”, ”oxidative stress”, ”epidemiology”, ”malaria” and ”aging”. Between 2016 and 2017, the most used keywords used by more than one institution were ”Alzheimer’s disease”, ”Parkinson’s disease”, ”cystic fibrosis”, ”nanoparticles”, ”climate change”, ”climate variability”, ”rheuma-

toid arthritis", "obesity", "epidemiology", "aging" and "quality of life". In the last time period in study, the most frequent keywords that appeared in the ranking of more than one institution were "Parkinson's disease", "ecosystem services", "climate change", "amyotrophic lateral sclerosis", "nanoparticles", "cytotoxicity", "obesity", "aging", "cancer", "exercise" and "quality of life".

Impact

As already observed, the impact of the research produced by the UL is influenced in a similar fashion by the production of each institution. Intuitively, the institutions whose production of documents is higher also present a higher likelihood of being cited. Regarding this metric, there is a group of highly cited institutions, constituted by IST and FCUL, whose publications' citations contribute to more than 77.0% of the citations of the UL, followed by a group of moderately cited institutions constituted by FMUL, iMM, FFUL, ISA and, in the last period, ISEG and finally a group of least cited institutions.

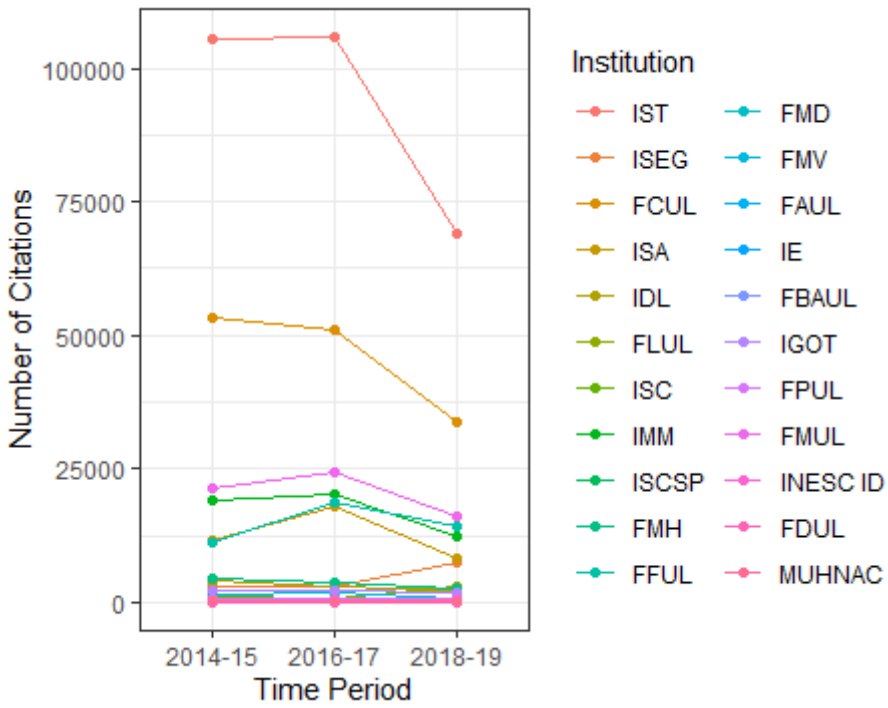

Figure 4.9: Evolution of the number of citations the publications, produced in each of the periods of time in study, of all domains of each institution of the UL acquire.

As can be seen in figure 4.10, the overall trend of the evolution of the number of citations acquired is descendent. However, half of the institutions (IST,ISA, iMM, ISCSP, FFUL, FMV, FBAUL, IGOT, FPUL, FMUL and FDUL) present a slight increase in the number of citations from the first to the second time period. The most surprising result is that ISEG, contrary to the overall trend, presents a very significant growth between 2016/17 and 2018/19.

In the health domain, the three most cited institutions, IST, FMUL and FCUL, contribute to more than

70.0% of the citations obtained by publications in the health domains of the UL. The evolution of the citations of each institutions' publications can be observed in figure 4.10. It can be observed that there is an overall trend of surge in the number of citations from the first to the second time period, followed by a decrease in the number of citations.

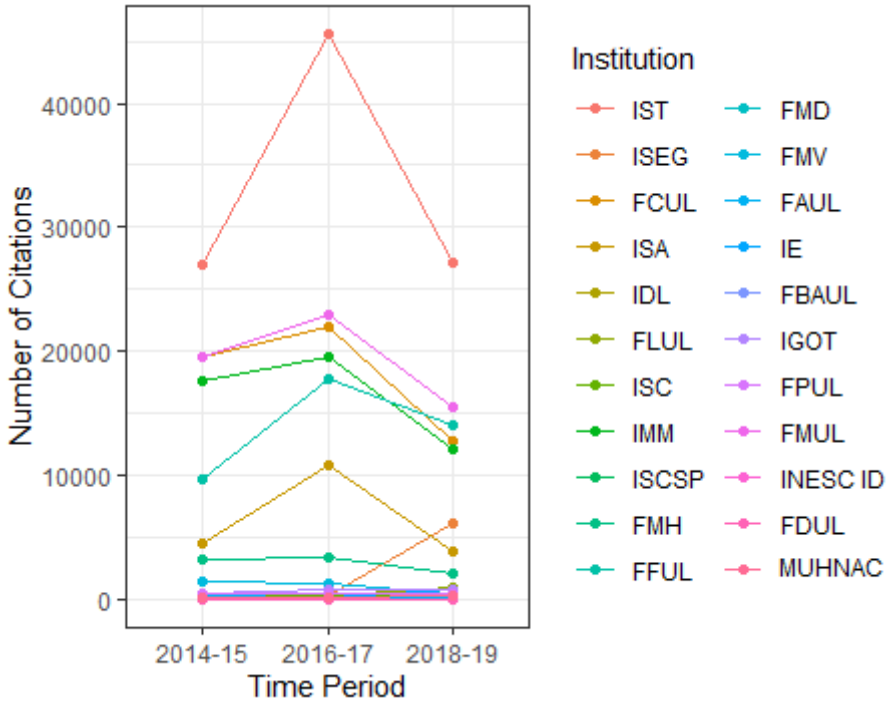

Figure 4.10: Evolution of the number of citations the publications, produced in each of the periods of time in study, of the health domain of each institution of the UL acquire.

Finally, the number of publications produced by an institution influences the number of citations received by that same institution. Having that in mind, the number of citations received by each institution has been normalized to the number of publications that institution has produced. In figure 4.11, this metric can be observed for both all publications of the UL and, in figure 4.12 the publications of the UL in the health domain.

For all publications, IGOT and iMM always present the highest value of average citations per year per document, except for the last time period in which ISEG presents the maximal value. The evolution of this metric is also of relevance. This value does not have a coherent tendency between all institutions. However, there are institutions that present a clear crescent tendency, i.e. this metric increases in the two transitions between time periods in study, in the average number of citations per document per year, namely, FAUL, FCUL, FFUL, FMH, FMUL, IDL, IE, IGOT, iMM, ISCSP, ISEG, IST and MUHNAC.

In the health domain, as can be witness when looking at the figure, IDL dominates the average citations per year per document in the first time period, but in the next time period it is surpassed by

IGOT, whom is then outpaced by ISEG. In the health domain, there are even less clear trends and only 50% of the institutions display an unambiguous growth tendency, namely, FBAUL, FCUL, FFUL, FMH, FMUL, IGOT, iMM, ISC, ISEG, IST and MUHNAC.

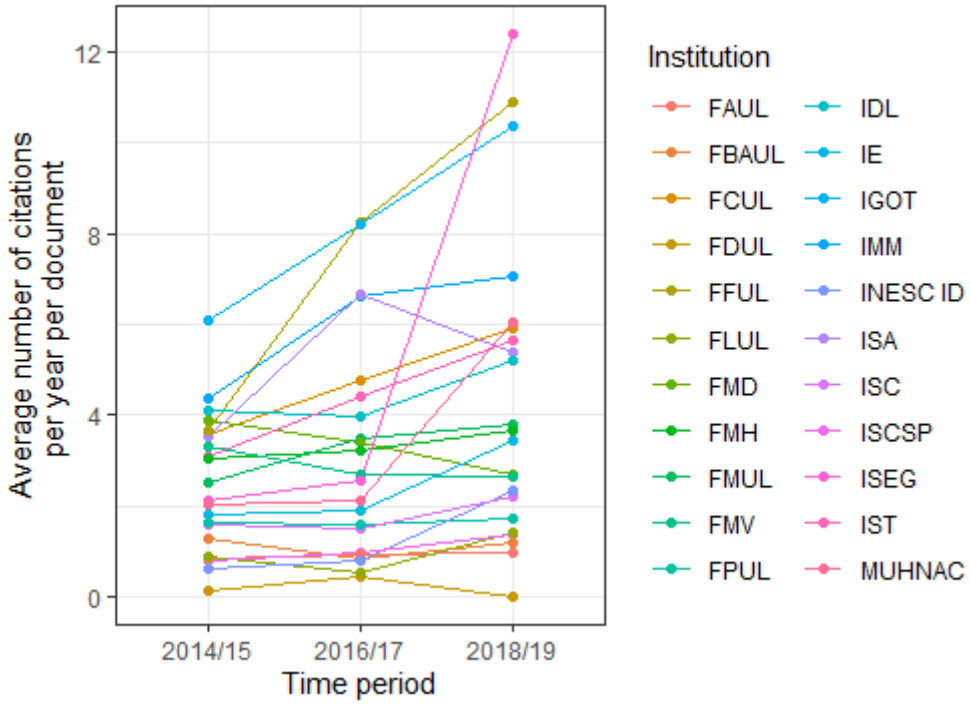

Figure 4.11: Evolution of the average number of citations per document produced per year all publications, produced in each of the periods of time in study, of each institution of the UL acquire.

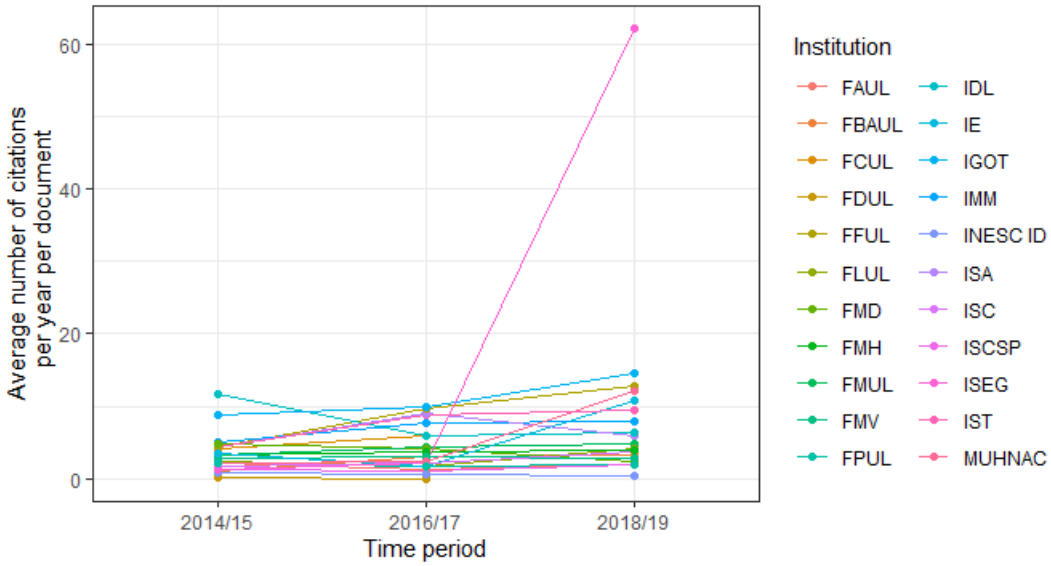

Figure 4.12: Evolution of the average number of citations per document produced per year the publications, produced in each of the periods of time in study, of the health domain of each institution of the UL acquire.

4.2 Collaboration and Keywords' Network - An analysis

To built a deeper understanding of both collaboration and the structure of the field that contribute to the health publications of the UL, social network analysis was used to build a collaboration network, based on co-authorship between institutions, and a keyword network, based on the co-occurrence between keywords associated by authors to their publications. In this section, those networks, as well as the metrics that characterize them, are shown.

4.2.1 Institutions' Network Data

The study of collaboration networks can provide information about the evolution of partnerships and aid in the management of collaborations. In that sense, the collaboration network, for each time period, in study of the publications in the health domain of the UL was created.

In figure 4.13, the collaboration network of the publications produced by the UL, in the health domain, from the first time period can be observed. The most influential institutions form five clusters where the volume of collaborations is higher. It can be noticed that there are clusters that present a superior collaboration density (red and yellow clusters), while there are clusters whose collaboration density is inferior and their institutions are placed more sparsely in the network (green, blue and purple clusters). In this network, the institutions of the UL are all located in the blue cluster indicating that they collaborate more closely with each other than with the institutions in the other clusters. The only node centrality measure that was possible to compute was degree centrality, which was only computed to the institutions of the UL. The institutions that presented a higher degree centrality was IST, which presented 7196 links, followed by FCUL with 5112 links and FMUL with 3848 links.

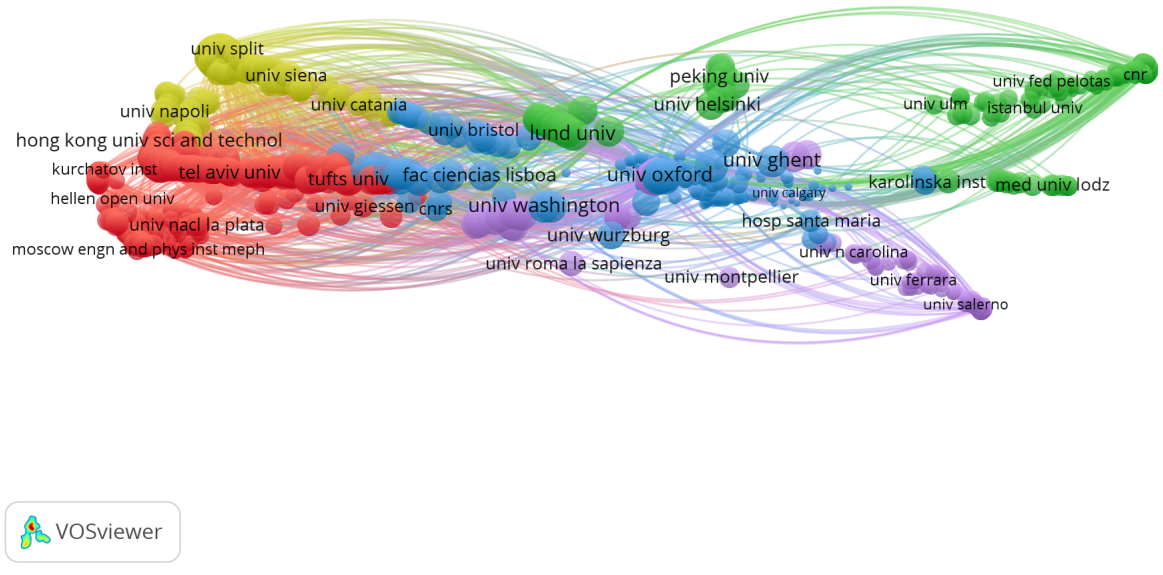

Figure 4.13: Collaboration network of publications produced by the UL classified as being of the health domain for the time period of 2014/15. Nodes represent institutions and links represent co-authorship of publications. In the figure, five clusters, defined by different colours, can be observed. Due to visualizations constraints only the one thousand most influential nodes are represented.

In figure 4.14, the collaboration network created by the co-authorship of publications of the health domain produced by the UL between 2016 and 2017 can be seen. The number of clusters formed by the most influential institutions decreased from 5 to 4. As had already happened in the first time period, there are clusters whose density of collaboration is superior (green and yellow cluster). However, in opposition to what happened in the first time period, the institutions of the UL are scattered in the network. IST, FCUL and ISA can still be found in the same cluster, the red one, but the FMUL is located in the blue cluster collaborating more closely with institutions of the health domain. iMM is found in the yellow cluster and FFUL in the green cluster. IST remains as the UL's institution with the highest degree centrality, participating in 8054 partnerships, followed again by FCUL that presents 5198 links and by FMUL with 4756 links.

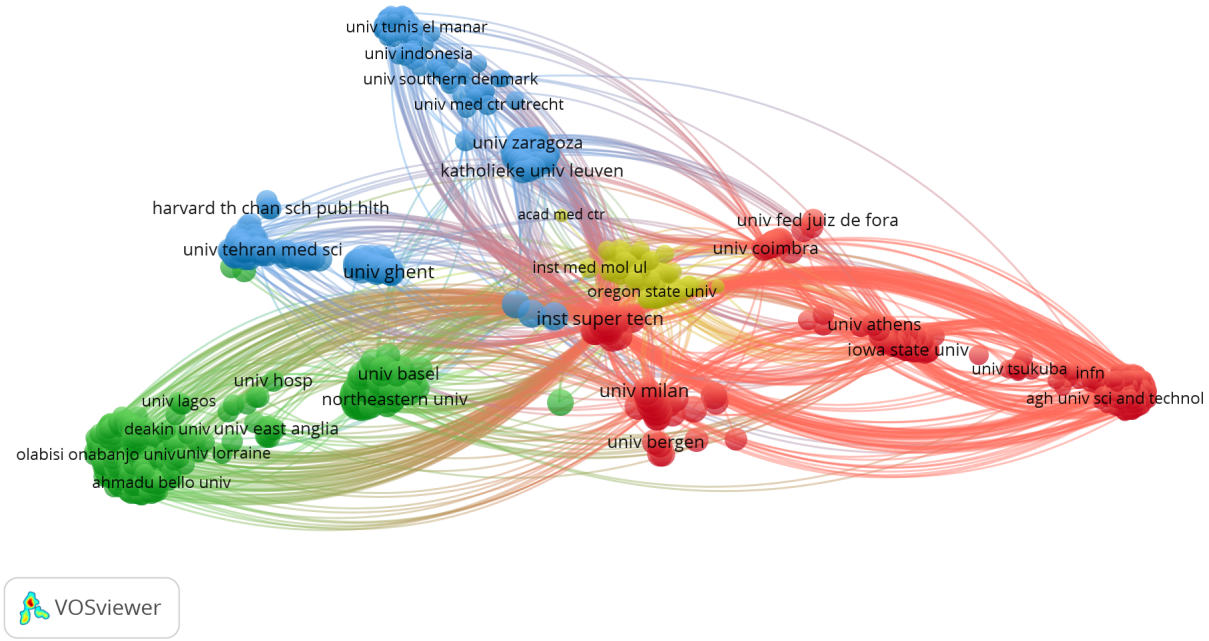

Figure 4.14: Collaboration network of publications produced by the UL classified as being of the health domain for the time period of 2016/17. Nodes represent institutions and links represent co-authorship of publications. In the figure, four clusters, defined by different colours, can be observed. Due to visualizations constraints only the one thousand most influential nodes are represented.

In figure 4.15, the collaboration network created by the 1000 most influential institutions that co-authored publications of the health domain produced with the UL between 2018 and 2019 is presented. Once again the number of clusters created by the institutions decreases to three and once more there is a densely packed cluster (in red) and two clusters whose collaboration is more sparsely located. The

most productive institutions of the UL are once more distributed between clusters. IST and FCUL can be found in the green cluster, while FMUL and iMM can be found in the blue cluster. Finally, FFUL is located in the red cluster. In this time period, FCUL surpasses IST as the most connected institution, with 13572 link, while IST presents 11156. Finally, the third most connected institution becomes iMM with 5780 links.

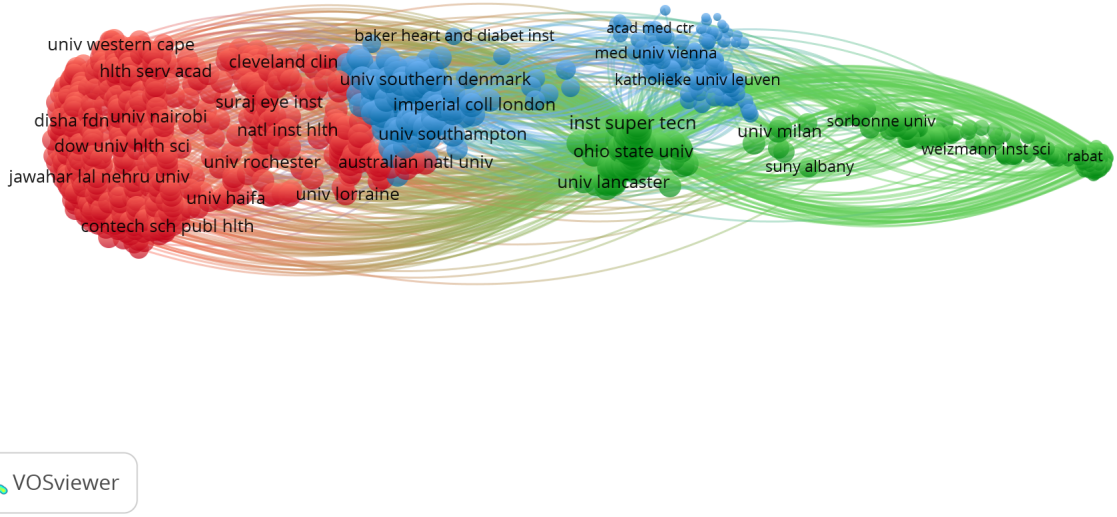

Figure 4.15: Collaboration network of publications produced by the UL classified as being of the health domain for the time period of 2018/19. Nodes represent institutions and links represent co-authorship of publications. In the figure, three clusters, defined by different colours, can be observed. Due to visualizations constraints only the one thousand most influential nodes are represented.

Besides understanding how institutions are positioned in the network and how their relative contribution influences the network, the network as a whole should also be looked upon. In table 4.8, the global metrics of the network can be found. As had already been demonstrated in the evolution of the number of institutions, the number of nodes increases, as well as the number of connections that increase by 330.6% from the first time period to the second and by 70.6% from the second to the third time period. Both the average path length and the diameter remain constant in the first transition, but decrease in the second time period. However, after a significant growth of the density of the network from 2014/15 to 2016/17, there is a relevant decline in the density from the second time period to the third. Finally, as the links increased at a faster pace than the number of nodes, a increase in the average degree can be

found.

Table 4.8: Collaboration network's metrics for publications in the health domain produced by the UL for the three time periods studied in this project. The unit of analysis were institutions, and as such, nodes represent institutions and the links represent co-authorship of publications.

	2014/15	2016/17	2018/19
Number of Nodes	4631	6928	11351
Number of Links	520348	2240634	3822370
Average Path Length	2.36	2.36	2.21
Diameter	5	5	4
Density	0.024	0.047	0.030
Average Degree	112.36	323.42	336.74

4.2.2 Keywords' Network Data

Besides understanding the research areas in which an institution publishes and the most used keywords, creating a network of co-occurrence of keywords may also provide important information regarding the cognitive field of an institution, as well as the evolution of innovation. The keywords' networks of the three time periods can be found in the appendix, in figures A.1, A.2 and A.3 and in table 4.9 the main metrics that characterize the networks can be found.

As already noted, there is an increase in the number of keywords used, as well as the number of links between them, which grew by 36.3%, followed by an increase of 26.5%. The average path length decreased in both transitions between time periods and the diameter, in spite of remaining constant between 2014/15 and 2016/17, declined in the third time period in analysis. Despite these contractions regarding the paths between nodes in the network, there is a shrinkage in the density of the network. As would be expected, as the number of nodes increases at a slower pace than the number of links, there is an increase in the average degree of the network. Finally, the number of clusters created by the most influential keywords presents a decreasing trend.

Table 4.9: Keywords' network's metrics for publications in the health domain produced by the UL for the three time periods studied in this project. The unit of analysis were the keywords authors associated to their publications, and as such, nodes represent keywords and the links represent co-occurrences of keywords.

	2014/15	2016/17	2018/19
Number of Nodes	9396	11951	14293
Number of Links	56132	76498	96792
Average Path Length	5.70	5.53	5.16
Diameter	14	14	12
Density	0.00064	0.00054	0.00046
Average Degree	5.97	6.40	6.77
Number of Clusters (1000 most influential keywords)	55	47	40

5

Discussion

Contents

Scientific progress has always been linked to improvements in society and how people live their lives, even more so in the past two centuries. The invention of the electric dynamo, computers, pasteurisation all contributed to societal developments. Nowadays, scientific research is more widespread and accessible to most of us than ever and with its growth and crescent availability also came the need to understand it and study it. The human being, specially the researcher, is obsessed with measuring, because it is through measuring that the understanding of the world may come. Having this in mind, the measurement of science soon became a need. In this thesis, the scientific production of the UL has been studied, with a special focus on the health domain. As all scientific studies, this one presents contributions and shortcomings and in this section the main ones identified are presented. After that, the results are discussed on the light of the objectives of the thesis and the previously read literature on the area.

5.1 Advantages of the methodology used

In this project bibliometric and SNA methods were used to accomplish the goals of describing the scientific production of the UL, as well as its evolution, both for all its publications and its publications on the health domain. Firstly, the main advantage of using bibliometrics to describe both productivity and impact of scientific research is related to the relative inexpensiveness of the method and the relatively fast learning curve needed to perform this kind of study.

In bibliometric and SNA, the indicators used to perform measurements and comparisons should have a meaning and should be thought of to fit the study to perform. There was no use of metrics which choice was advised against. For example, related to the bibliometric metrics, neither the h-index, which, as already mentioned, reduces two very important indicators into one, nor the JIF, which use is advised in evaluating scientific journals were used. In addition to that, due to the sensitivity to outliers of mean values, whenever it was possible the median and distribution of values were used.

Furthermore, bibliometric metrics should always be applied from a benchmarking or evolutionary perspective, to assign the indicators some meaning [12]. In this work, this recommendation was followed. Indicators were always looked upon from a comparative perspective, between all publications and the publications of the health domain, or from an evolutionary perspective, by focusing on the progression of the indicators from time period to time period.

The recommendations for levels of aggregation were also followed. Looking at higher levels of aggregation, i.e. departments and universities, instead of lower, i.e. researchers and research groups, increase the reliability of bibliometric results.

The use of the WoS allows for an easy acquisition of the data and due to the existence of R-packages like Bibliometrix, the process is easily reproducible. This is an inexpensive and reproducible process that

can be performed for any level of aggregation to which data is available.

In assessing the health domain, not only publications present on PubMed or from medicine journals were taken into account. Using the methodology described in section 2.3, it was possible to gather publications of the many diverse fields that publish findings that may belong or influence the health domain. Having that in mind, this study was exhaustively multidisciplinary and provided information on a variety of areas that contribute to health research.

Furthermore, regarding the science mapping part, Bibliometrix, the SNA package, and VOSviewer make it very straightforward to create bibliometric networks, as well as to obtain the metrics that characterize them and to visualize them. The visualization of the networks provide an intuitive way to look at the data and to both inform policy makers and researchers as well as to appeal to the information of the general population.

Additionally, regarding the measuring of innovation, to the best of the knowledge of the author of this thesis, there are no studies that use keyword co-occurrence mapping and SNA to evaluate innovation of scientific research in the context of universities. The method to perform this part of the study was adapted from methods used in innovation research in firms and as such may be a new tool that may be useful to assess the progress of innovation in academic research. However, it should be studied more thoroughly to be considered as such.

Finally, this project may contribute to the discussion of strategies for research and collaboration between researchers of the UL, not to mention support the constant characterization and assessment of the scientific production of this institution.

5.2 Disadvantages of the methodology used

Despite presenting diverse advantages, this methodology, like every methodology, is not without its shortcomings. Firstly, regarding the use of bibliometrics as a way to measure science, relying solely on bibliometric measures to inform decision making or to compare institutions may increase the adoption of malpractices regarding production of scientific research or citation of that same research [24]. Having this in mind, it is very important to remind the reader that bibliometric analysis should never be exclusively be used to inform policy makers. Besides that, regarding this study in particular, no normalization was performed for any indicator, regarding field differences or institutions' characteristics (e.g. funding acquired, number of researchers, etc.). Citations or publication counts were never normalized for area of research, or number of researchers investigating in each institution, or funding. Because of that, this study cannot be used to directly compare the institutions of the UL with each other.

As already mentioned, scientific production, citations counts and co-authorship are only proxies for productivity, impact and collaboration. Scientific production and citations may not be the only way that

universities foster scientific research, as promoting conferences, the transference of knowledge and education also push for scientific advancements. Additionally, joint supervision of Master or pHD's thesis, joint teachership or joint organization of conferences or events are further collaborative activities that are not taken into account when assessing collaboration through co-authorship of publications.

Regarding the data retrieval, in spite of WoS being extremely useful, intuitive and providing many bibliographical information, its coverage is not very strong for the Social Sciences and Humanities and there is a negative bias towards the indexation of non-english speaking journals. Thus, this work may also be biased against these areas and against the institutions that publish in portuguese-speaking journals.

Even though WoS provides an organization identifier that provides a better standardization of the institutions' names when compared with other available databases, it is still not ideal. There was the need to standardize the name of the institutions in this work. However, this process was extremely exhaustive and there is the possibility of it not being complete, which may partially influence the results. In addition to this, the selection of publications from the health domain was only 85% accurate and as such there may be a slight misrepresentation of the publications of this field.

Besides that, due to the lack of computational power, the acquisition of all the metrics needed to perform network analysis was not possible, such as betweenness or closeness centrality. Hence, there is a part of the portrait of collaboration in the UL's network that is still hidden.

Finally, due to a lack of time and resources, it was not possible to clean spelling mistakes or variations of keywords, nor to aggregate them in groups. The picture that is obtained from the keywords' analysis may be very crude and needs to be worked upon.

5.3 Discussion of the results for the University of Lisbon (UL)

In this work, the description and analysis of several dimensions of scientific research were investigated. Firstly, productivity was looked upon, followed by content of the research, as well as impact, collaboration and innovation.

Using publication counts as a measure of productivity, it can be concluded that general productivity in the UL is increasing. It can also be noticed that the productivity of the health domain in the UL is increasing at a faster pace. This productivity increase may be explained by several factors. Firstly, it may be the case that the UL is seeing a growth in the human resources it employs. Besides that, some institutions or departments may have seen an increase in the funding received or that certain more productive scientific fields have seen an increase in investment [7]. Finally, as already mentioned an increase in collaboration, which is verified can have a positive impact on productivity [7, 9, 37–39].

It can also be concluded that the most productive research areas of the UL are Technology and

Physical Sciences. However, their relative contribution is decreasing, while the relative contribution of the Social Sciences is increasing. In the health domain, Medicine and Life Sciences lead in relative contribution and the contribution of the Social Sciences to this field is also modestly increasing.

Using keywords' uniqueness and the knowledge about the keywords' co-occurrence network, it can be concluded that the diversity topics is growing and the inter-changeability between topics also present a slight sign of growth. The number of unique keywords surges between every period of study, which may indicate an increase in exploratory innovation [79,80], i.e. researchers are broadening the cognitive field of the research in the health domain produced by the UL. Besides that, there is an increase in the number of links between the nodes, above the increase in the number of nodes, complemented with a decrease in the average path length, diameter and number of clusters, which may signal an investment in exploitative innovation and in multidisciplinarity [78–80]. This variation in innovation may also be explained by the evolution of the collaboration network. Collaboration with new institutions or countries may influence the network of knowledge elements as new expertise may be brought to play [80]. These results are antagonized by a decrease in the density of the network from time period to time period and thus should be complemented with other data.

Regarding the impact of the research of the UL, there are no clear conclusions to come to, since there may not have been enough time for the publications produced in the last two time periods in study to acquire citations. However, it is interesting to understand that most publications do not acquire a large number of citations, while some outliers acquire a lot, and this is seen both for the generality of publications and for the health domain. Besides that, on average, publications of the health domain acquire more citations than the average of all publications of the UL. The fact that the publications produced in the second time period in both all fields and in the health domain, and the publications produces in the third period of the health domain already have a higher average number of citations per document per year may indicate that the publications produced in these time periods will over time acquire more citations than the ones produced in the first time period, but there is no certainty in this. This may be explained by a stronger investment in the communication of scientific findings [12] or an increase of collaboration [43,44], that is observed in this study. Finally, another important factor to have in mind is that around 20% of the publications of the UL of the time periods in study are yet to be cited.

Regarding collaboration, there is an increase in the number of authors, institutions and participating countries from the first time period to the second and third, both for all publications and for the publications of the health domain. Using co-authorship as a proxy for collaboration, it can be assessed that collaboration in the publications of the UL has been increasing. The fact that in the health domain, the countries that collaborate the most with the UL are located more closely may reveal the importance of geographical proximity for the development of health research. On another note, it could also mean that there are more opportunities for funding for collaborations between EU countries or a preference

for similar cultural beliefs, this is something that should be studied. Furthermore, when analysing the collaboration networks and comparing its evolution, their size has been increasing, as well as the number of links between institutions, which supports the hypothesis of an increase in collaboration. The fact that the average path length, diameter and number of clusters are declining further corroborates the plausibility of an surging in collaboration. However, the decrease in the density, between the second and third time period, contradicts the absolute affirmation that collaboration is increasing. If both networks are visualized carefully, there seems to be a considerable increase in the density of some clusters but an even more significant decline in the density of other clusters, which may explain the slight decrease in this metric. The density slight decrease from the second to the third time period does not seem to be enough to overcome an overall surge in collaboration. This increase in collaboration, both domestic and international, may explain the increase in productivity [9, 37–39] and in the impact, here translated as average citations per document per year, of the health domain of the UL [43, 44]. Finally, starting in the second time period there seems to be a shift from collaboration between institutions' of the UL to external collaboration, this hypothesis is based on the crescent dispersion of these institutions by the different clusters of the collaboration network.

A brief analysis of the scientific production of the institutions' of the UL was also performed. In this study, it was found out that for all the fields of study there is a set of very productive institutions, followed by one of moderately productive institutions. In the health domain, this distinction is less heterogeneous and there is a group of six scientific institutions whose production contributes highly to the production of scientific knowledge in the health domain of the UL. These distributions in the number of publications may indicate that the Lotka's law of productivity is applicable to the institutions of the UL. However, more data would have to be analysed and a more thorough study devised.

The most used keywords associated by authors to their publications common to more than one institution were also presented because they may demonstrate some opportunities of collaboration. In the health domain, there seem to be potential for collaboration specially in subtopics such as "Parkinson's disease", "Alzheimer's disease", "obesity", "quality of life" and "climate change". There is the possibility that these keywords are already the most frequent because they result from the collaborations between the institutions, this has to be further investigated.

Finally, the impact of each institution was also calculated and plotted. Impact follows a distribution similar to the productivity and there are few institutions that acquire many citations and many institutions that acquire less citations. It could be interesting to evaluate if this distribution follows the Pareto's law. Besides the count of the absolute number of citations, this value was normalized for the number of documents produced by each institution and for the number of years in which the documents could acquire citations, painting a clearer picture of how the impact of each institution relates to each other. It is interesting to notice that the most productive institutions, which are also the ones that contribute the most to

the overall impact achieved by the UL, do not have a high expression in the average number of citations per document per year, neither for all publications nor for the health domain. This may be related to several factors that influence the acquisition of citations by publications. It may be the case that institutions like IGOT, FFUL and iMM are overperforming the most productive institutions in the communication they make of their findings [12]. It may also be the case that these institutions are participating in more fruitful collaborations with diverse institutions [43] or even international collaborations [44]. Finally, it could also be the case that the most productive institutions, in the scientific culture of ”publish or perish”, end up publishing many documents on topics that are already exhausted [12] or that do not yet have interest from the scientific community.

6

Conclusion, Recommendations and Future Perspectives

Contents

6.1 Concluding Remarks

Scientific progress is at the heart of societal development. To build the world of tomorrow, a better, improved version of the one there is today, a more equitable, just and fair, a world that promotes equal opportunities for all, safety, peace, a world in which no child suffers from hunger or diseases that could easily be treated, a world in which each human being can advocate for their own health and well being, the role of science is going to be vital. To improve the making of science and scientific research, there needs to be a study of this area. It was in that context that this thesis emerged. Its goals, as already discussed, were to describe the scientific research produced by the UL, particularly in the health domain, as well as to study collaboration and innovation in this domain. It is hoped that these description and analysis may inform policy making in the UL to contribute to an amelioration of research in this institution and at a national level.

To accomplish the goals presented in the previous paragraph firstly, there was the need to come acquaintance with the world of bibliometrics and SNA, through a bibliographic review. After that, the methodology was built grounded on the literature in the area and the objectives to accomplish. It was required to retrieve the data regarding the publications that the UL produced between 2014 and 2019. The retrieval of the data was performed using the bibliographical database WoS. After the retrieval of the data, there was the need to standardize the data, particularly the affiliations' data. These data were then analysed and visualized. The main findings that were arrived using this process are summarized in the next paragraphs.

Firstly, it was found that productivity, as measured by publication counts, in the UL both for all the publications and for the publications in the health domain has increased in the studied years. This growth can be explained by several already mentioned factors, such as an increase in the human resources or funding or possibly due to the increase in collaboration.

Besides that, it was noted that for all publications, the Research Area that contributes the most for the publications produced by the UL was Technology, which may result from the fact that the most productive institution of the UL is a school of engineering and technology, IST. In the health domain, as it would be expected, the area that contributes the most to the publications of the UL is medicine, which is also in line with the most productive institution, for the second and third time period in study, FMUL. In both domains, the contribution of the Social Sciences has been increasing an it may be interesting to watch out for the future of this research area in the UL.

When analysing the most frequent keywords associated to publications of the UL, it is interesting to see that health research is gaining prominence, as ”Parkinson's disease” and ”cancer” enter the top-10 of most used keywords. Furthermore, it can be noticed that there is a growing interest in the topic of climate change.

Regarding impact in the UL, it can be highlighted that on average publications in the health domain

tend to have a higher impact than the average of all publications. Besides that, in the health domain the superior average number of citations per document per year in the second and third time period relative to the first time period in study may indicate that on the long run, publications of these time periods will have a higher impact than the ones produced in the first time period.

Concerning collaboration, there seem to be several aspects that support the hypothesis that collaboration in publications of the UL is increasing. The number of authors, institutions and countries participating in those publications. Additionally, in the health domain, the analysis of the institutional collaboration networks seem to further back the plausibility of this hypothesis.

Finally, in a very limited study of innovation, the evolution of the number of unique keywords and of the keywords co-occurrence networks may imply that innovation in the health domain in the UL is increasing.

6.2 Recommendations

As already denoted, this study is not without its limitations and there is a lot the author of this thesis has learned since the beginning of this thesis. As such, a few recommendations both to inform policy making and the future production of similar studies.

Firstly, having in mind the results and discussion of this thesis, it is very important to remind the reader that this study consists above all on a description of the scientific production. It is not an evaluative or comparative study neither of the UL, nor constituting institutions. The results should be looked upon the context of each institution. Furthermore, the sole use of bibliometric data to inform policy making should be avoided and there should be a refrain from supporting the ”publish or perish” science culture.

Developing from the work performed in this study, it would be relevant to carry out studies like this with more regularity, not only for the UL as a whole, but for each institution, for its departments, research groups and even individual researchers, to aid them in improving their scientific process. For this, the standardization of the affiliations and researchers' names in the publications produced is of utmost importance. Additionally, a manual to ground future bibliometric studies of the UL should be developed.

Finally, due to the impact that collaboration may have on science productivity, impact and innovation it may be relevant to foster and encourage researchers from the UL to attend inter-institutional meetings of researchers who investigate in the health domain.

6.3 Future Work

The shortcomings of the methodology used have already been presented, the future work is largely based on those shortcomings and on aspects whose analysis was not possible to perform in this study.

Firstly, it would be important to improve the method to retrieve publications of the health domain, in order to obtain more accurate results. Furthermore, with the aid of data or computer science, the methods to standardize the affiliations' names should be made more rigorous. Besides that, in a future study, there should also be a pre-processing of the keywords to depict a more clear picture of knowledge domains and areas of study. Additionally, a way to overcome the biases that arose in the social sciences and the arts and humanities should be found.

The description of the collaboration and keywords co-occurrence networks is incomplete. Finding the metrics of individual nodes is of importance to understand how the network evolves and which actors promote communication the most. As such, a future study should be performed to evaluate betweenness and closeness centrality.

To further deepen this analysis, there are several hypothesis that should be studied. Studying the influence of collaboration on productivity, impact and innovation of the research produced by the UL, studying how the collaboration network of researchers and institutions influences the knowledge network or cognitive structure of the fields investigated by the UL and studying what is the contribution of the cognitive field of each institution on the cognitive field of the UL are all important to understand what policies should be put into place.

Hereafter, a study of collaboration that goes beyond scientific production should be performed.

All the recommendations and suggestions made in this section can be promising and may contribute to the development of scientific research in the UL and consequently at our scale be the drop in the ocean of science that fights for a better future.

Bibliography

[1] M. Thelwall, L. Vaughan, and L. Björneborn, "Webometrics," *Annual review of information science and technology*, vol. 39, no. 1, pp. 81–135, 2005.

[2] N. Harwood, "An interview-based study of the functions of citations in academic writing across two disciplines," *Journal of pragmatics*, vol. 41, no. 3, pp. 497–518, 2009.

[3] H. Perkin, "History of universities," in *International handbook of higher education*. Springer, 2007, pp. 159–205.

[4] D. Wang and A.-L. Barabási, *The science of science*. Cambridge University Press, 2021.

[5] E. Leahey, "From sole investigator to team scientist: Trends in the practice and study of research collaboration," *Annual review of sociology*, vol. 42, pp. 81–100, 2016.

[6] B. F. Jones, S. Wuchty, and B. Uzzi, "Multi-university research teams: Shifting impact, geography, and stratification in science," *science*, vol. 322, no. 5905, pp. 1259–1262, 2008.

[7] S. Lee and B. Bozeman, "The impact of research collaboration on scientific productivity," *Social studies of science*, vol. 35, no. 5, pp. 673–702, 2005.

[8] S. Wuchty, B. F. Jones, and B. Uzzi, "The increasing dominance of teams in production of knowledge," *Science*, vol. 316, no. 5827, pp. 1036–1039, 2007.

[9] G. Abramo, A. C. D'Angelo, and G. Murgia, "The relationship among research productivity, research collaboration, and their determinants," *Journal of Informetrics*, vol. 11, no. 4, pp. 1016–1030, 2017.

[10] B. Godin, "On the origins of bibliometrics," *Scientometrics*, vol. 68, no. 1, pp. 109–133, 2006.

[11] A. Ramos and C. S. Sarrico, "Past performance does not guarantee future results: lessons from the evaluation of research units in portugal," *Research Evaluation*, vol. 25, no. 1, pp. 94–106, 2016.

[12] H. F. Moed, W. Burger, J. Frankfort, and A. F. Van Raan, "The use of bibliometric data for the measurement of university research performance," *Research policy*, vol. 14, no. 3, pp. 131–149, 1985.

[13] P. Kokol, H. Blažun Vošner, and J. Završnik, "Application of bibliometrics in medicine: a historical bibliometrics analysis," *Health Information & Libraries Journal*, vol. 38, no. 2, pp. 125–138, 2021.

[14] D. F. Thompson and C. K. Walker, "A descriptive and historical review of bibliometrics with applications to medical sciences," *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, vol. 35, no. 6, pp. 551–559, 2015.

[15] P. Kokol, J. Završnik, and H. B. Vošner, "Bibliographic-based identification of hot future research topics: an opportunity for hospital librarianship," *Journal of Hospital Librarianship*, vol. 18, no. 4, pp. 315–322, 2018.

[16] R. da Univerdidade de Lisboa, "Ulisboa em números," January 2018.

[17] "Bibliometria," https://www.ulisboa.pt/info/bibliometria, accessed: 2021-10-26.

[18] "Redesaúde," https://www.ulisboa.pt/info/redesaude, accessed: 2021-05-05.

[19] E. Garfield, "" science citation index"-a new dimension in indexing," *Science*, vol. 144, no. 3619, pp. 649–654, 1964.

[20] D. Jacobs, "Demystification of bibliometrics, scientometrics, informetrics and webometrics," in *11th DIS Annual Conference*, 2010, pp. 1–19.

[21] W. W. Hood and C. S. Wilson, "The literature of bibliometrics, scientometrics, and informetrics," *Scientometrics*, vol. 52, no. 2, pp. 291–314, 2001.

[22] A. Cavadas, "Visualising the collaboration network of a european marine research infrastructure: A bibliometric and social network analysis," *U. Porto Journal of Engineering*, vol. 6, no. 2, pp. 98–118, 2020.

[23] A. F. Van Raan, "Measuring science," in *Handbook of quantitative science and technology research*. Springer, 2004, pp. 19–50.

[24] S. Haustein and V. Larivière, "The use of bibliometrics for assessing research: Possibilities, limitations and adverse effects," in *Incentives and performance*. Springer, 2015, pp. 121–139.

[25] A. Agarwal, D. Durairajanayagam, S. Tatagari, S. C. Esteves, A. Harlev, R. Henkel, S. Roychoudhury, S. Homa, N. G. Puchalt, R. Ramasamy *et al.*, "Bibliometrics: tracking research impact by selecting the appropriate metrics," *Asian journal of andrology*, vol. 18, no. 2, p. 296, 2016.

[26] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma, and F. Herrera, "An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field," *Journal of informetrics*, vol. 5, no. 1, pp. 146–166, 2011.

[27] C. R. Sugimoto and V. Larivière, *Measuring research: what everyone needs to know*. Oxford University Press, 2018.

[28] W. Glänzel, H. F. Moed, U. Schmoch, and M. Thelwall, *Springer handbook of science and technology indicators*. Springer Nature, 2019.

[29] J. S. Katz and B. R. Martin, "What is research collaboration?" *Research policy*, vol. 26, no. 1, pp. 1–18, 1997.

[30] E. Garfield, "Journal impact factor: a brief review," *Cmaj*, vol. 161, no. 8, pp. 979–980, 1999.

[31] ——, "The evolution of the science citation index," *International microbiology*, vol. 10, no. 1, p. 65, 2007.

[32] J. Scott, "Social network analysis," *Sociology*, vol. 22, no. 1, pp. 109–127, 1988.

[33] R. Pranckutė, "Web of science (wos) and scopus: The titans of bibliographic information in today's academic world," *Publications*, vol. 9, no. 1, p. 12, 2021.

[34] P. D. Allison and J. S. Long, "Departmental effects on scientific productivity," *American sociological review*, pp. 469–478, 1990.

[35] J. S. Long and R. McGinnis, "Organizational context and scientific productivity," *American sociological review*, pp. 422–442, 1981.

[36] B. A. Jacob and L. Lefgren, "The impact of research grant funding on scientific productivity," *Journal of public economics*, vol. 95, no. 9-10, pp. 1168–1177, 2011.

[37] S. E. Hampton and J. N. Parker, "Collaboration and productivity in scientific synthesis," *BioScience*, vol. 61, no. 11, pp. 900–910, 2011.

[38] G. Abramo, C. A. D'Angelo, and F. Di Costa, "Research collaboration and productivity: is there correlation?" *Higher education*, vol. 57, no. 2, pp. 155–171, 2009.

[39] J. A. Castillo and M. A. Powell, "Research productivity and international collaboration: a study of ecuadorian science," *Journal of Hispanic Higher Education*, vol. 19, no. 4, pp. 369–387, 2020.

[40] I. Tahamtan and L. Bornmann, "What do citation counts measure? an updated review of studies on citations in scientific documents published between 2006 and 2018," *Scientometrics*, vol. 121, no. 3, pp. 1635–1684, 2019.

[41] B. Wang, Y. Bu, and Y. Xu, "A quantitative exploration on reasons for citing articles from the perspective of cited authors," *Scientometrics*, vol. 116, no. 2, pp. 675–687, 2018.

[42] C. W. Belter, "Bibliometric indicators: opportunities and limits," *Journal of the Medical Library Association: JMLA*, vol. 103, no. 4, p. 219, 2015.

[43] M. Franceschet and A. Costantini, "The effect of scholar collaboration on impact and quality of academic papers," *Journal of informetrics*, vol. 4, no. 4, pp. 540–553, 2010.

[44] T. V. Nguyen, T. P. Ho-Le, and U. V. Le, "International collaboration in scientific research in vietnam: an analysis of patterns and impact," *Scientometrics*, vol. 110, no. 2, pp. 1035–1051, 2017.

[45] L. Bornmann, R. Haunschild, and R. Mutz, "Should citations be field-normalized in evaluative bibliometrics? an empirical analysis based on propensity score matching," *Journal of Informetrics*, vol. 14, no. 4, p. 101098, 2020.

[46] T. Kiesslich, M. Beyreis, G. Zimmermann, and A. Traweger, "Citation inequality and the journal impact factor: median, mean,(does it) matter?" *Scientometrics*, vol. 126, no. 2, pp. 1249–1269, 2021.

[47] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National academy of Sciences*, vol. 102, no. 46, pp. 16 569–16 572, 2005.

[48] L. Waltman and N. J. Van Eck, "The inconsistency of the h-index," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 2, pp. 406–415, 2012.

[49] N. J. Van Eck and L. Waltman, "Visualizing bibliometric networks," in *Measuring scholarly impact*. Springer, 2014, pp. 285–320.

[50] A. Van Raan, "Advances in bibliometric analysis: research performance assessment and science mapping," *Bibliometrics Use and Abuse in the Review of Research Performance*, vol. 87, pp. 17–28, 2014.

[51] C. Chen, R. Dubin, and T. Schultz, "Science mapping," in *Encyclopedia of Information Science and Technology, Third Edition*. IGI Global, 2015, pp. 4171–4184.

[52] C. Chen, "Science mapping: a systematic review of the literature," *Journal of data and information science*, vol. 2, no. 2, 2017.

[53] I. Zupic and T. Čater, "Bibliometric methods in management and organization," *Organizational research methods*, vol. 18, no. 3, pp. 429–472, 2015.

[54] R. N. Broadus, "Early approaches to bibliometrics," *Journal of the American Society for Information Science*, vol. 38, no. 2, pp. 127–129, 1987.

[55] M. E. Falagas, E. I. Pitsouni, G. A. Malietzis, and G. Pappas, “Comparison of pubmed, scopus, web of science, and google scholar: strengths and weaknesses,” *The FASEB journal*, vol. 22, no. 2, pp. 338–342, 2008.

[56] K. Canese and S. Weis, “Pubmed: the bibliographic database,” *The NCBI Handbook*, vol. 2, p. 1, 2013.

[57] H. Bukvova, “Studying research collaboration: A literature review,” 2010.

[58] P. W. Mattessich and B. R. Monsey, *Collaboration: what makes it work. A review of research literature on factors influencing successful collaboration.* ERIC, 1992.

[59] D. H. Sonnenwald, “Scientific collaboration,” *Annual review of information science and technology*, vol. 41, no. 1, pp. 643–681, 2007.

[60] M. F. Fox and C. A. Faver, “Independence and cooperation in research: The motivations and costs of collaboration,” *The Journal of Higher Education*, vol. 55, no. 3, pp. 347–359, 1984.

[61] D. Beaver, “Reflections on scientific collaboration (and its study): past, present, and future,” *Scientometrics*, vol. 52, no. 3, pp. 365–377, 2001.

[62] V. A. Haines, J. Godley, and P. Hawe, “Understanding interdisciplinary collaborations as social networks,” *American journal of community psychology*, vol. 47, no. 1-2, pp. 1–11, 2011.

[63] B. d. P. F. e Fonseca, R. B. Sampaio, M. V. de Araújo Fonseca, and F. Zicker, “Co-authorship network analysis in health research: method and potential use,” *Health research policy and systems*, vol. 14, no. 1, pp. 1–10, 2016.

[64] B. Bozeman, D. Fay, and C. P. Slade, “Research collaboration in universities and academic entrepreneurship: the-state-of-the-art,” *The journal of technology transfer*, vol. 38, no. 1, pp. 1–67, 2013.

[65] I. Chompalov and W. Shrum, “Institutional collaboration in science: A typology of technological practice,” *Science, Technology, & Human Values*, vol. 24, no. 3, pp. 338–372, 1999.

[66] L. D. Sargent and L. E. Waters, “Careers and academic research collaborations: An inductive process framework for understanding successful collaborations,” *Journal of Vocational Behavior*, vol. 64, no. 2, pp. 308–319, 2004.

[67] K. Subramanyam, “Bibliometric studies of research collaboration: A review,” *Journal of information Science*, vol. 6, no. 1, pp. 33–38, 1983.

[68] B. Ponomariov and C. Boardman, "What is co-authorship?" *Scientometrics*, vol. 109, no. 3, pp. 1939–1963, 2016.

[69] A.-L. Barabási, "Network science," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1987, p. 20120375, 2013.

[70] J. L. Gross and J. Yellen, *Handbook of graph theory*. CRC press, 2003.

[71] M. E. Newman, "Coauthorship networks and patterns of scientific collaboration," *Proceedings of the national academy of sciences*, vol. 101, no. suppl 1, pp. 5200–5205, 2004.

[72] S. W. Aboelela, J. A. Merrill, K. M. Carley, and E. Larson, "Social network analysis to evaluate an interdisciplinary research center." *Journal of Research Administration*, vol. 38, no. 1, pp. 61–75, 2007.

[73] J. Godley, G. Barron, and A. M. Sharma, "Using social network analysis to assess collaboration in health research," *Journal of Healthcare, Science & the Humanities*, vol. 1, no. 2, pp. 99–116, 2011.

[74] D. Knoke and S. Yang, *Social network analysis*. Sage Publications, 2019.

[75] B. P. Fonseca, E. Fernandes, and M. V. Fonseca, "Collaboration in science and technology organizations of the public sector: A network perspective," *Science and Public Policy*, vol. 44, no. 1, pp. 37–49, 2017.

[76] H. Etzkowitz and L. Leydesdorff, "The triple helix as a model for innovation," *Science and public policy*, vol. 25, no. 3, pp. 195–203, 1998.

[77] B. H. Hall and N. Rosenberg, *Handbook of the Economics of Innovation*. Elsevier, 2010, vol. 1.

[78] M. L. Weitzman, "Recombinant growth," *The Quarterly Journal of Economics*, vol. 113, no. 2, pp. 331–360, 1998.

[79] G. Carnabuci and J. Bruggeman, "Knowledge specialization, knowledge brokerage and the uneven growth of technology domains," *Social forces*, vol. 88, no. 2, pp. 607–641, 2009.

[80] C. Wang, S. Rodan, M. Fruin, and X. Xu, "Knowledge networks, collaboration networks, and exploratory innovation," *Academy of Management Journal*, vol. 57, no. 2, pp. 484–514, 2014.

[81] J. Guan and N. Liu, "Exploitative and exploratory innovations in knowledge network and collaboration network: A patent analysis in the technological field of nano-energy," *Research policy*, vol. 45, no. 1, pp. 97–112, 2016.

[82] H.-N. Su and P.-C. Lee, "Mapping knowledge structure by keyword co-occurrence: a first look at journal papers in technology foresight," *Scientometrics*, vol. 85, no. 1, pp. 65–79, 2010.

[83] R. da Universidade de Lisboa, “Escolas — ulisboa,” https://www.ulisboa.pt/escolas, September 2021.

[84] M. Aria and C. Cuccurullo, “bibliometrix: An r-tool for comprehensive science mapping analysis,” *Journal of Informetrics*, vol. 11, no. 4, pp. 959–975, 2017. [Online]. Available: https://doi.org/10.1016/j.joi.2017.08.007

[85] S. Guidelines Review Committee, R. Health, and Research, “Who guideline on self-care interventions for health and well-being,” July 2021.

[86] A. . D. Data, “In focus: 2021,” September 2021.

[87] A. to Assistive Technology, A. t. M. Medical Devices, H. P. P. Health Products, M. D. Standards, and Diagnostics, “Who compendium of innovative health technologies for low-resource settings 2021. covid-19 and other health priorities,” August 2021.

[88] J. P. Kandhasamy and S. Balamurali, “Performance analysis of classifier models to predict diabetes mellitus,” *Procedia Computer Science*, vol. 47, pp. 45–51, 2015.

[89] M. Aria and C. Cuccurullo, “bibliometrix: An r-tool for comprehensive science mapping analysis,” *Journal of Informetrics*, vol. 11, no. 4, pp. 959–975, 2017. [Online]. Available: https://doi.org/10.1016/j.joi.2017.08.007

[90] C. T. Butts, *sna: Tools for Social Network Analysis*, 2020, r package version 2.6. [Online]. Available: https://CRAN.R-project.org/package=sna

[91] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. [Online]. Available: https://ggplot2.tidyverse.org

[92] N. J. Van Eck and L. Waltman, “Vosviewer manual,” *Leiden: Univeristeit Leiden*, vol. 1, no. 1, pp. 1–53, 2013.

[93] “Research areas (categories / classification),” https://images.webofknowledge.com/images/help/WOS/hp_research_areas_easca.html, accessed: 2021-10-10.

[94] “Vosviewer,” https://www.vosviewer.com/, accessed: 2021-4-10.

Appendix

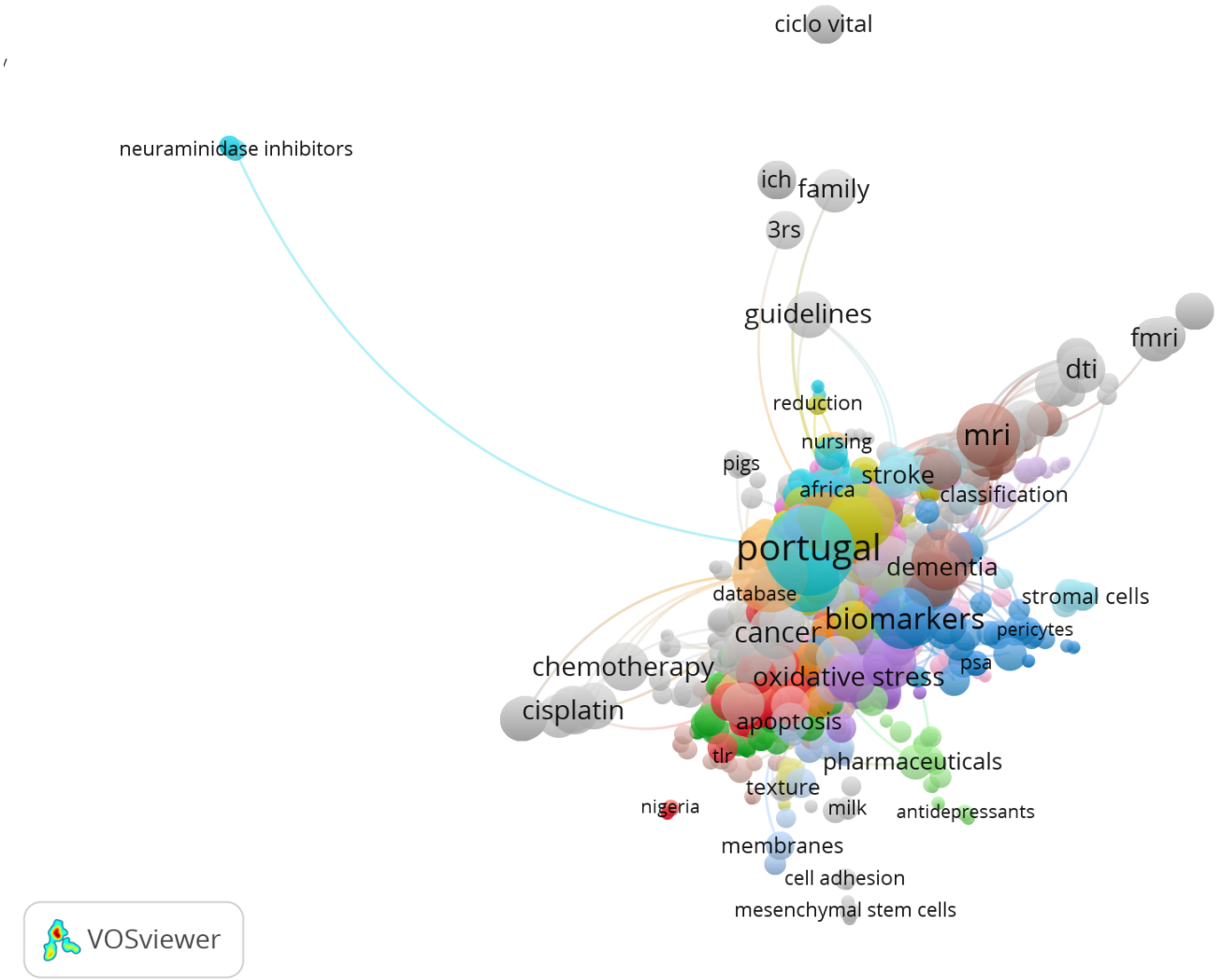

Figure A.1: Keywords' co-occurrence network of the publications of the health domain produced by the UL between 2014 and 2015. Nodes represent keywords and links represent a co-occurrence of the keywords it connects. Due to visualization constraints only the one thousand most used keywords are used.

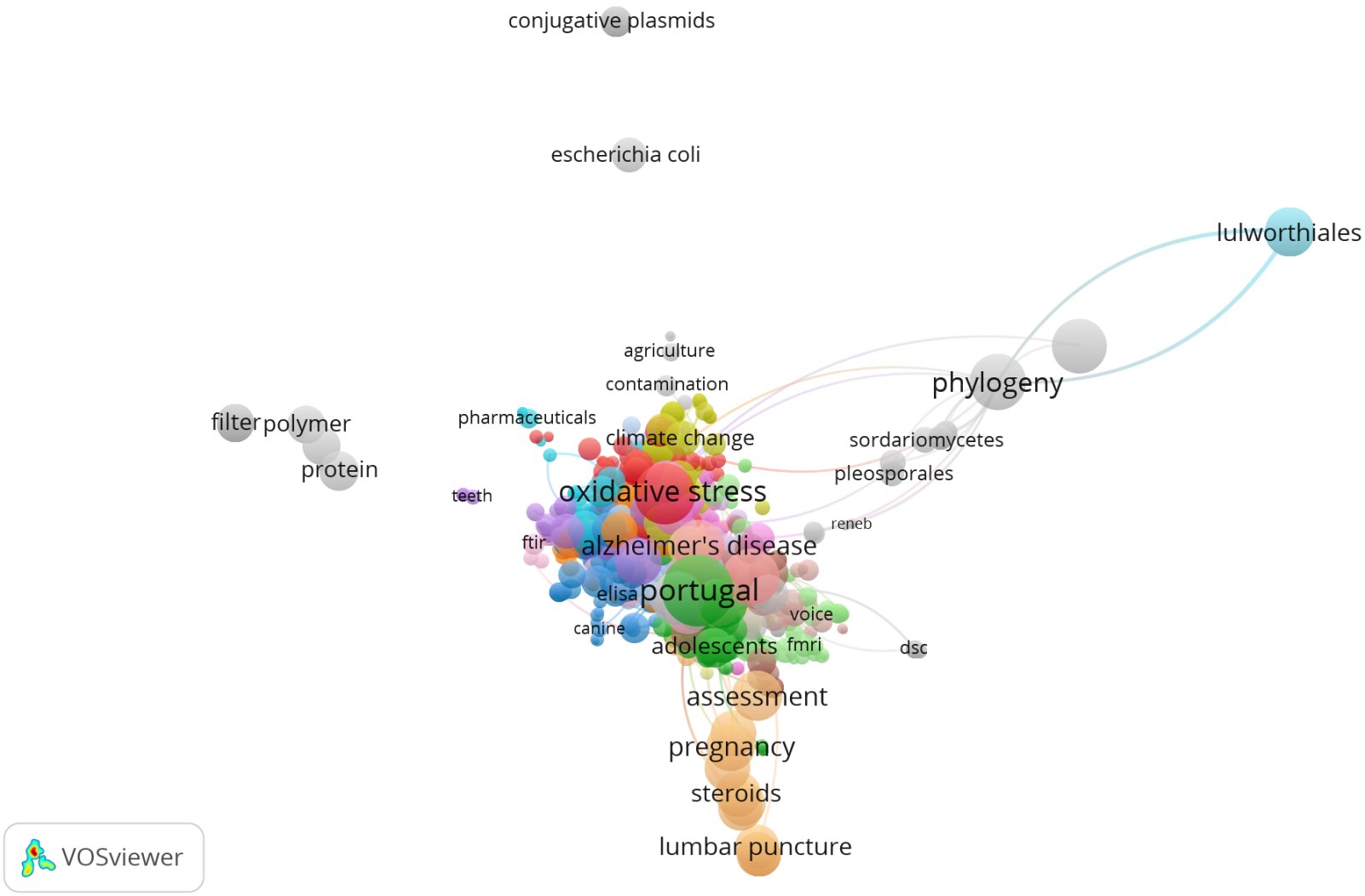

Figure A.2: Keywords' co-occurrence network of the publications of the health domain produced by the UL between 2016 and 2017. Nodes represent keywords and links represent a co-occurrence of the keywords it connects. Due to visualization constraints only the one thousand most used keywords are used.

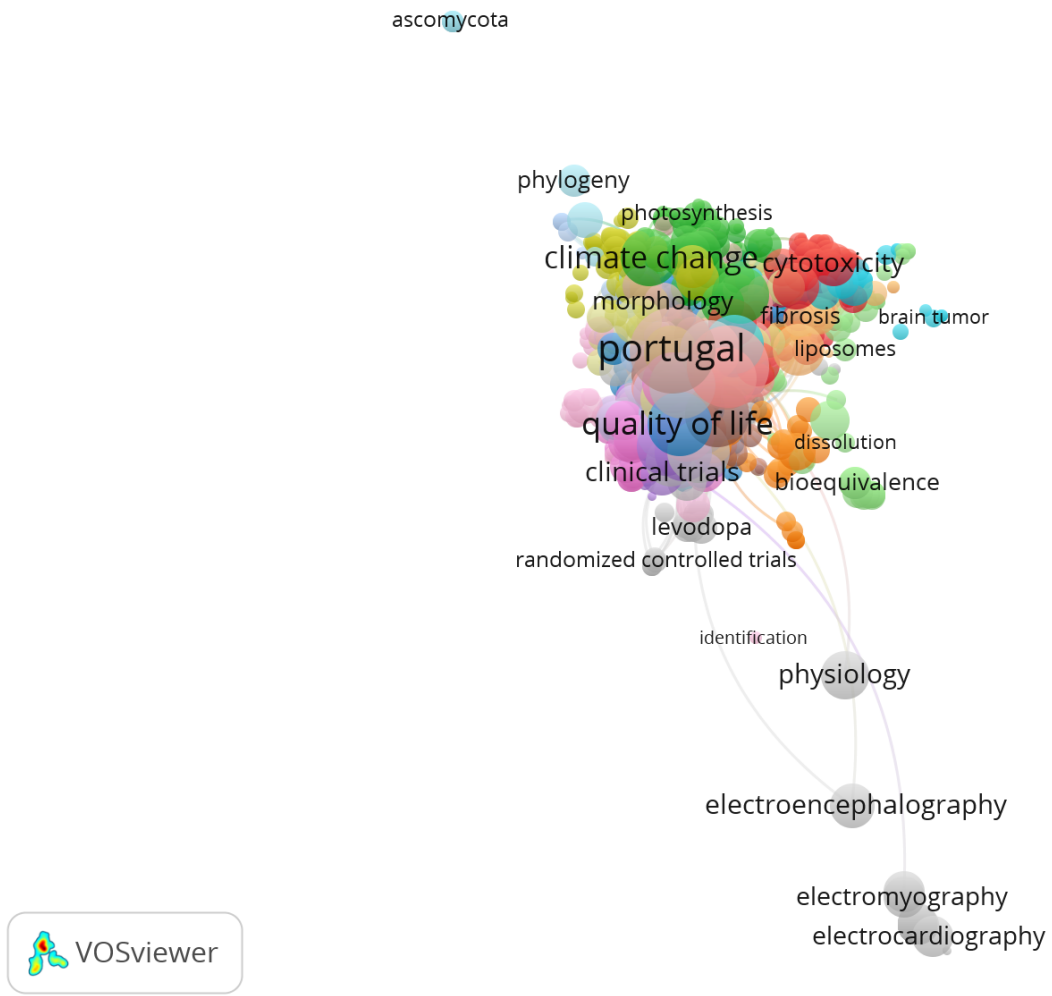

Figure A.3: Keywords' co-occurrence network of the publications of the health domain produced by the UL between 2018 and 2019. Nodes represent keywords and links represent a co-occurrence of the keywords it connects. Due to visualization constraints only the one thousand most used keywords are used.