Universiteit Leiden
The Netherlands

Opleiding Informatica

Team Configuration & Creativity: *Ability, Specialisation, Diversification and Knowledge Overlap*

for a Bachelor Thesis

Arjan Visser
s1698087@umail.leidenuniv.nl

Supervisor: J. Wang [hw18] Second Reader: You-Na Lee
Supervisor:j.wang@sbb.leidenuniv.nl

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

28/06/2018

Abstract

What kind of team configuration benefits creativity the most? This thesis explores effects of member ability composition, specialisation vs. diversification, and knowledge overlap between members. Ability gives insight into the capability of the individual team members, diversification captures the amount of information embodied in individuals, and knowledge overlap is about the cognitive distance between team members. We focus on the biomedical discipline and consequently build a dataset from PubMed and the Databank of Illinois. For each publication in the targeted year its forward citations are retrieved as a proxy for team creativity. For each author, all the MeSH terms used his/her prior publications are retrieved, as well as citations received by his/her prior publications. MeSH terms are then used for measuring individual diversification and pairwise knowledge overlap, and the prior citations for individual ability. We run an OLS regression for explaining citations and find that a homogeneous team with specialised individuals has the largest number of citations.

Keywords: **I**nformatica & Economie; **D**iversification; **G**eneralisation; **S**pecialisation; **P**ubMed; **ME**Dline; **S**cience Teams; **S**cientists; **K**nowledge frontier; **I**nvisible College; **C**reativity; **OL**S

Acknowledgements

This research would not have been able to conduct, if Torvik, Vetle I. and Smalheiser, Neil R. wouldn't have lend the Author-ity 2009 - PubMed author name disambiguated dataset for unique author identifiers [TS18]. Furthermore this research would have much less quality if it wasn't for the guidance and advise of the supervisor. [hw18]

Contents

Abstract 2

Acknowledgements 3

Contents 4

List of Tables 6

List of Figures 7

1 Introduction 1

2 Theoretical Framework 3

2.1 Science as a Collaborative Effort . . . 4

2.1.1 Creativity . . . 4

2.1.2 General Discussion on Collaborations . . . 4

2.1.2.1 Knowledge Burden . . . 6

2.1.2.2 Islands of Automation . . . 7

2.1.3 Specialisation . . . 8

2.1.4 Diversification . . . 9

2.1.5 Diversification versus Specialisation . . . 11

2.1.6 Heterogeneous and Homogeneous . . . 11

2.2 Problem Statement . . . 12

2.3 Team Configuration . . . 13

2.3.1 Member abilities . . . 13

2.3.2 Specialist or Generalist . . . 14

2.3.3 Knowledge Overlap . . . 15

3 Research Design 18

3.1 Data . . . 18

3.2 Variables . . . 22

3.2.1 Depended . . . 22

3.2.2 Independent . . . 22

3.2.2.1 Ability and Skill . . . 22

3.2.2.2 Generalist or Specialist . . . 23

3.2.2.3 Knowledge Overlap . . . 23

3.2.3 Control . . . 23

3.3 Methodology . . . 25

4 Results 27

4.1 Descriptive . . . 27

4.2 Predictive . . . 30
4.2.1 Regressor . . . 30
4.2.2 Member abilities and skills . . . 32
4.2.3 Specialist or generalist . . . 32
4.2.4 Knowledge overlap . . . 33
4.2.5 Control Variables . . . 34

5 Conclusions **35**

Bibliography **38**

A Appendix **41**

List of Tables

3.1 Variables in the master dataset explained 24

4.1 Description of the master log1p table 28
4.2 Main results 31

A.1 Robustness, no control for meshcount paper 44
A.2 All fieldnames which were used as dummy variables 50

List of Figures

2.1 A Venn diagram of the knowledge frontier 7
2.2 The marked area is the N personality, the yellow a generalist, the blue a specialist, a part of the yellow combined with the grey the T personality 10
2.3 The theoretical framework dimension in a matrix format 13
2.4 The model of the theory ⇢= negative relation; →= positive relation 16
2.5 The theoretical framework dimension in a matrix format, filled with the conceptual variables. A – indicates a low amount, whereas the + indicates a high amount 17

3.1 A part of the MeSH hierarchical database [Inf14] 19
3.2 Histograms of the independent variables for the log1p master dataset 21

4.1 Correlation heatmap based on the non-parametric Spearman measure for log1p dataset 29

A.1 Correlation heatmap based on the non-parametric kendall measure for log1p dataset 41
A.2 Correlation heatmap based on the non-parametric spearman measure for log1p dataset 42
A.3 Correlation heatmap based on the pearson r measure for log1p dataset 43

CHAPTER 1

INTRODUCTION

The story about the discovery of gravity is common knowledge. Isaac Newton was sitting in the garden, when an apple fell on his head. However, if gravity was discovered in a modern setting than the apple would have fallen on: *'Xin Lu, Jan Jansen, Hans Wolf Bach, Michelle Harrington and Isaac Newton'* during a collaboration effort. This is because of, as Benjamin F. Jones puts it, "...the death of the Renaissance man."- [Jon09] in the academic world.

A scientist can no longer be the all-round expert as was common in the past, but is forced to specialise if he wants to bring something to the table. [Jon09] [WJU07] Even when this scientist specialises, he can only hope to increase the smallest atomic value in terms of scientific knowledge. This due to the ever expanding *knowledge frontier* and complexity within science. [APV16] [Jon09] [Lea16] Thus, a collaboration is formed of multiple specialists, which have enough accumulated knowledge to produce significant research. [Jon09] However, overspecialisation is of concern as well, since this would lead to tunnel vision and a decrease in novelty and usefulness of the creative product. [Ver16] [Lea07] Which begs the question how a team of researches must balance specialisation and diversification for the individual level, plus homogeneous and heterogeneous for the team level. Both levels refer to either a focus on some categories or a broad spectrum of categories. Do note that the word balance is used and not *'chose between'*, since this would assume mutual exclusion.

The main goal of a scientific community is to push forward the knowledge frontier. [Ver16] Which only expands when new knowledge and understanding is added to the collective memory of the community. [WVS16] Such knowledge is driven if not caused, by novel and useful work. Thus, one must ask to which degree one must be specialised or generalized to increase the chance on novelty work. The main ingredient for novelty in any kind is human creativity. One is thus actually interested in the optimal distribution of specialisation and generalisation for the best creativity impact. This research investigates:

> What kind of team configuration in terms of ability, diversification/specialisation and knowledge overlap benefits creativity the most?

There have been investigations to answer this question, however few are empirical. There also has been research assuming mutual exclusion of specialisation and generalisation, while investigating its effects on an individual scientist level. This research doesn't assume mutual exclusion. On the contrary this research views a collaboration as interaction of individuals, thus a team configuration, where one can be diversified and specialised in certain degree. This is the first theoretical contribution.

The second contribution is by adding empirical research, while examining the overlap between team members' expertise directly, which were studied rather indirectly in previous team studies. This gives a fairly better insight into team configurations. Consequently, in a practical manner, this research would provide critical information for policy makers and money lenders. They could create policies favouring research teams, which have the highest change at creating creative products. Which is beneficial for the entire scientific community. Some might claim that it helps improve the world overall, because of that.

This research uses the following structure, in order to investigate this matter. First definitions are introduced together with related work in the theoretical framework. After which the research design is discussed. Thirdly, the results are presented and discussed in the same section. This research finalises with a conclusion and recommendations for future work.

Chapter 2

Theoretical Framework

The number of co-authored articles is increasing as the sole scientist model is shrinking in prevalence. [APV16] [Jon09] Wutchy et al [WJU07] supports this finding, he reports a plus minus 90% increase in scientific team sizes in 2007. These collaborative efforts aren't only formed in one respective field of science, cross collaboration has been come more common. [NKV17]

With collaborating comes new opportunities arise for specialising and diversifying. Previously, in the sole scientist model, one only had two options. A scientist could either keep learning in broad spectrum of fields effectively diversifying himself, or he could focus on one (sub)field and thus becoming a specialist. [Jon09] With a collaboration one can obtain a focused or broad spectrum of knowledge on a team level as well. One can create a team of only diversified scientists, or a team of specialists specialised in different knowledge categories, both have a broad spectrum of knowledge, but different dynamics. One could also have a team of specialists all specialised on the same categories, creating a focused knowledge spectrum. In conclusion one can mix similar and dissimilar team members. On a side note, this is applicable on other concepts then knowledge as well. This illustrates the power of team configuration.

As was discussed with a collaboration comes new opportunities. One can specialise and diversify on two levels, the individual level and the team level. On a individual level team members can have a certain degree of specialisation and a certain degree of diversification. [NKV17] On a team level one can have similar and dissimilar team members. [SFE15] [Ver16] [WTG15a] Each level has its own distinct effects on the team's end result. [Ver16] This is evident in that team level specialisation and/or diversification is no substitute for the individual level. The individual level leads to better results. [Ver16] The team level, on the other hand, is the most practical. One of the reasons why it is more practical has to do with the complexity of some fields or the deep specialisation that is necessary to grasp and fully understand a field. [LR08] Moreover many have argued that a team is more than the sum of its parts. [APV16] Apart from the atomic value an individual would bring a team, thus a mere addition, another team member would also add discussions, knowledge sharing and perhaps social benefits or even social costs.

As a result the discussion as to which team configuration on both levels would yield the best creative product is still ongoing. [LR08] [LBS16] There is logically no T-model answer, which applicable for each scientific discipline, philosophy is quite different from computer science. To help this discussion, this research explores the effects of member ability composition, specialisation vs. diversification, and knowledge overlap between members. These variables are conceptually defined in the Team Configuration section. Before this section comes to play the relevant

literature is summarised in the section Science as a Collaborative Effort

2.1 Science as a Collaborative Effort

Academic investigation has become a team effort. [AGT16] Depending on what definition of team one uses; one can claim there was never a sole scientist model. Some papers have argued that advice or aide in more mundane tasks should already be enough reason to speak of a team effort. [Lea16] For example, a Renaissance scientist relying on advice of the spouse would be for some enough reason to speak of a collaborative effort. A stricter definition, which admittedly fails to describe the context, [Lea16] [SFE15] would be all registered co-authors. This stricter definition is used in this paper. There is a problem with this measure apart from contextual one. It has been come more common to register PhD students, supervisors or quite random individuals who were previously only acknowledged as a co-author [AGT16]. This results in some level of noise in late entries in databases. Which can influence measurements.

2.1.1 Creativity

No matter which definition for a team one uses, the goal remains the same. The goal of a scientific inquiry is to obtain knowledge. [SFE15] Obtaining knowledge, which is already known has little value. Therefore, one wants to obtain knowledge which wasn‘t known before or got forgotten. A very important aspect of this is novel and useful work. Both novelty and usefulness can be summarized by how creative a work is. [LWW14] Therefore the goal of a scientist or a collaborative effort is to create creative works.

Creativity is one of the more abstract terms, to which many interpretations are credited. [LWW14] Social sciences describe creativity as thinking differently, having a different motivation or multiple different experiences to pull from. This is however hard to measure and is arguable. It is defined by the Cambridge dictionary as:*“the ability to produce original and unusual ideas, or to make something new or imaginative“* [CN17]. Some scientific works have transformed this to the ability of combining pieces of knowledge in a novel way. [Ama83] However, [AZM09] rightfully expands this idea. He states that some discoveries are indeed incremental and can be found by creatively combining existing knowledge, however, other discoveries will require innovative and radical methods to lead to new novel work. This research follows the definition from Amabile [Ama83] and expands it with the comment from Manso [AZM09]. Creativity is thus the novel recombination of knowledge pieces in a useful way and creating novel approaches to achieve non-incremental useful and novel products.

Note how novelty and usefulness are important dimensions of creativity. [LWW14] Novelty or originality is an important aspect of creativity. However, novelty in its own is not enough. One can have the most novel idea, but if it’s too far ahead of its time, or blatantly not useful, then it would not lead to new knowledge or insights, since it cannot be used. The same applies for usefulness. One can plagiarise creatively in such a way that it goes unnoticed, but it is not novel and thus introduces no new knowledge, effectively eluding the scientific goal. Therefore, novelty and usefulness are both important aspects of creativity which cannot be taken apart. Thus, creativity is about making novel and useful products.

2.1.2 General Discussion on Collaborations

As of now the goal of a collaboration is stated. This goal doesn‘t explain why the sole scientist model is so good as gone. After all they have the same goal. Therefore the advantages of a collaboration are explained. This will give us insight in too the benfit/cost trade off that scientists have made through the years. After the advatages the main cited drivers are discussed. This should give insight in to which forces are responsible for the switch of a sole

scientist model to team efforts.

A collaboration has some general advantages and drawbacks that effects creativity. Let's start by kicking in an open door. A collaboration has multiple scientists. Generally one tries to further specialise or diversify knowledge categories in a collaboration, such that the problem statement can efficiently be investigated. When one diversifies on a team level, dissimilar authors are introduced in to a team. When this is the case one speaks of a heterogeneous team in terms of knowledge. The homogeneous team in terms of knowledge is of course it's opposite extreme. Notice that more authors are introduced, thus the team size increases. Recent empirical studies have found that the mean of team sizes is doubling each year. [WJU07] Walsh et al [LWW14] have investigated what the effect of team size is. From which they founded a positive relation with citation count. Moreover they found a U-inverted relationship between team size and novelty. They argue that the relationship with novelty, has to do with the attempt to diversify. [LWW14] [Lea16] Which confirms with what we stated earlier. This implies that a diversified knowledge set benefits novelty, which is correct. Dissimilar authors introduce more knowledge pieces from different domains. With more unique knowledge pieces a team can solve the puzzle of the research question in a more ways than one. More importantly it increases innovative potential. All with all more information pieces is beneficial for the creative product.
Furthermore, more team members, increases resources, these extra resources makes teams far more able to handle the unforeseen, which improves overall quality. [LWW14] This improves the potential usefulness of the work. There is however, a downside to team size as well, these large teams made it necessary to adopt organisational like behaviour, such as management and policy setting, which brings costs. [LWW14] [LR08] The managerial costs are paid with team resources, which could have been used for the research itself. This results in that teams have less resources available for proper research, which in turn means that less potential problem statements can be investigated. Moreover more limited resources may forbid the use of more radical approaches, which decreases creative potential.

Another advantage has to do with a common claim that a team is more than the sum of its parts. [APV16] One of the concepts that supports this claim is that of knowledge sharing. [APV16] [NKV17] Scientist share knowledge of their expertise with other members of the team. As a result other team members enter new knowledge fields. This is added value for the individual. Moreover the setting of learning something new, leads to asking questions and discussion. This potentially increases innovation through new perspectives, which is beneficial for the creative product. However, when such a team gets a big win, then they are less likely to go bold in the future, because of path dependency. [NKV17] This means that a radical team becomes less radical overtime and with that the chances of creative work decreases. Other research has shown that scientists rather go into new fields with people from their own social network, rather than doing it alone. [SFE15] While diversified individuals may have a social network spanning multiple fields, a specialist has commonly only a network spanning its home field. This creates a snowball effect generalists are more inclined to further diversify and specialist are more inclined to specialise due to their social networks. [NKV17] Moreover another research suggests that generalist rather cooperate with other generalist, rather than specialist. Which increases the snowball effect. [Lea16]

There are other benefits and drawbacks of teams as well, for the sake of completeness these are summed up: [Lea16] [LWW14] [APV16] [LBS16] [WJU07]

1. Increased productivity
2. Increased grant likelihood
3. Better visibility
4. Higher impacting work
5. Better quality

6. Synergy possibilities

However, there are drawback as well:

1. Los on communication and coordination
2. Misaligned goals
3. Lacking in field language
4. Administration cost
5. Increased emotional cost and pressure
6. Free riding

Recall that teams behave like organisations nowadays. In this light, this paper introduces the term 'management and agency cost' relative to the financial world, as the cost made to compensate for free riding, goal misalignment, mental state of team members and administration costs.

While benefits and drawbacks to collaboration may give insight into the benefit cost trade off of scientists, it fails to explain the rapid growth of collaborative efforts in science. Therefore the most common cited drivers of collaborations are discussed. The first driver of a collaboration is that of costs and shared resources. [WJU07] By having multiple shoulders carry the same weight, the burden becomes lighter to carry. This gives opportunity to conduct more expensive research. Thus new problem statements can be solved in a creative way. Furthermore radical approaches which were first costly are now fair game, which increases the creative potential of the product. A reason closely intertwined with this is that of competitive grants. [dRP18] By hoarding fellow scientists, one's opinion as to what to research next, becomes stronger and might be just enough to win a grant for which other researchers are competing. Besides winning the grant, having multiple views from other scientists incorporated in to the research proposal increases output quality, which in turn benefits usefulness. There are other collaboration drivers such as local policies, academic rank and demographic properties. The foremost reason for a collaboration is that of specialisation and/or diversification. [Jon09] [AGT16] This can be explained with modern problem statements as they are cross-field and increasingly ill defined. This forces collaboration in order to gain the expertise and knowledge to tackle the problem statement. [Lea16] Two other concepts are more to 'blame' for the importance of specialisation and diversification, these are the knowledge burden and the islands of automation.

2.1.2.1 Knowledge Burden

The reason that specialisation would lead to an increase in collaboration has to do with the knowledge burden. The knowledge burden is at the frontier of the drastic increase in the number of collaborations. [Jon09] The knowledge burden has to do with the gigantism of the so-called knowledge frontier. The knowledge frontier is the bleeding edge of the collective memory of science. [Jon09] [dRP18] Everything in the frontier is known, while everything outside is all that needs the be discovered or rediscovered, since knowledge can be forgotten as well. The goal of a research should be to push forward the knowledge frontier. [WVS16] Recall that this goal is for a great part achieved by creative work. [LWW14] However creative research is no guarantee for advancement of the knowledge frontier and advancement of the knowledge frontier is not necessarily a creative work. In this figure 2.1 a crude visualisation of the knowledge frontier can be found.

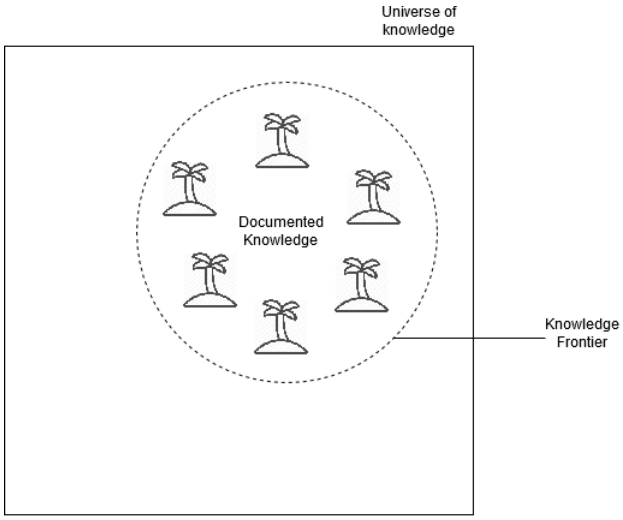


Figure 2.1: A Venn diagram of the knowledge frontier

The amount of documented knowledge available to a researcher is too large for a human being to learn and understand within a single lifetime. [APV16] To uncover new understandings and carry the 'knowledge frontier' one step further, a researcher must specialise within a discipline. This concept is called the knowledge burden. [Jon09] The effect of the knowledge burden is only enhanced by the modern 'eureka effect', which makes that new developments come at a fast rate, making it impossible to have in depth up-to-date knowledge in a broad scientific field. [Lea16] The increase in complexity strengthens this effect. [LR08] As a result of the increased specialisation of scientists, due to the knowledge burden, this poses a problem for the sole scientist model. It has become harder to comprehend a field or to effectively address a scientific question. This forces a scientist to collaborate in order to solve problem statements. Furthermore scientific fields are broad and highly productive. Consequently collaboration are necessary to gain the ability to match the fields high productivity, recall a familiar phrase: 'publish or perish'. [AGT16] [Lea16] Besides the productive ability the collaboration also needs to take place to accumulate enough knowledge to produce relevant work. [LWW14] This is in line with the knowledge burden. The knowledge burden gives a scientist two options, either learn more or specialise into a niche science. [Jon09] The learning more option decreases time for research immensely, so much so that almost everyone has chosen to specialise to a certain degree.

2.1.2.2 Islands of Automation

By specialising a researcher becomes a part of an active science sub community or discipline. [Lea16] [LR08] Thus, in practice the whole science community is a body consisting of limbs in form of disciplines. [BC73] While some disciplines overlap and actively collaborate with each other, most disciplines are inclined to no form of interaction. One could speak of a form of *'Islands of automation'* in science. These small island communities of knowledge specialists are alienated from each other, as a result effective communication comes at a cost The natural state of a discipline is inclined to no cross communication, due the resultant differentiation in specialising choices. [BC73] [Lea16] On a side note the idea of island automation in science 2.1.2.1 has overlap with the concept of the invisible college introduced by Diana Crane [BC73].

Continuing the concept of island automation in science. The actual meaning of island automation is first explained. Island automation is a term hailing from the system architecture discipline. 'Island of automation' are groups of systems consisting of one or more hard or software entities, who can't, won't or hardly can communicate with each other, while in-group communication is fine. This can lead to data duplication, wrong information, high costs and bad decision making. The same can be applied in social academic setting. Where the groups can be viewed as

scientific disciplines and the entities as scientists and academic communication systems. [TBM97] [NL99]

The are some reasons why disciplines are not inclined to cross communication. First is a social factor that affects all humans, a human prefers interaction of its own kind, with their own distinct customs. [LR08] Thus, it comes down to an English saying:" better is the devil you know, then the one you don't know". One rather suffers that what he knows, than that what is unknown. As a result scientists are known mostly in their own fields and therefore are even more inclined to collaborate only with each other. [ftAoS13]

Secondly pure specialisation yields higher productivity, visibility and salary. [LR08] [Lea07] [AGT16] Such that niche research specialisation is very attractive. [LR08] As a result the incentives to specialise may outweigh the incentives to diversify. While this beneficial for productivity and visibility, it lacks innovative potential. [Lea16]

Lastly some fields suffer from an affliction called over-specialisation. The result of which is that one is not interested in or can even understand the works in the foreign fields. Consequently one is only up-to-date in his own field. This causes a void in author networks, such that new experiences in the form of collaboration become rare. [LR08] [Lea07]. This results in less entries into new fields, through collaboration. Moreover one doesn‘t recieve other perspectives and doesn‘t receive incentives to debate. This results in a loss of innovative potential.

There are more reasons for this phenomenon, but they go beyond the relevance of this paper. To name one would be practical limitations, one rather cooperates with anyone in the same office building than someone that is in another country. One should note that these practical costs have been reduced due to modern technology. [APV16]

More importantly is the increased understanding of the value of cross islands collaboration and the battle against the inclination of not communicating. Scientists try to overcome this island structure by creating bridges between the islands through the means of collaboration and individual level diversification. By having multiple specialists of different fields in one collaboration, bridges are formed between different domains of knowledge. Other configurations are of course also possible, such that more information is introduced for a collaboration to use. This increase in information benefits problem solving. Moreover it is appreciated for its highly novel and useful character. While this team level diversification is no substitute for the individual level, which is still more beneficial. [Ver16] It is more practical for most scientists, which makes collaborating the most efficient way to meet the knowledge requirement for a modern problem statement.

To summarize specialisation has two distinct consequences on collaboration. First, disciplines are broad and are characterized as highly productive. [Lea16] Thus, under the concept of the knowledge burden it is hard for an individual scientist to obtain all relevant literature in a discipline, thus encouraging collaboration inside the discipline. [Lea16] Secondly cross islands collaboration is increasingly appreciated for its useful and novel character. Where collaborating is the most practical and efficient way to achieve cross domain knowledge, by introducing dissimilar members.

2.1.3 Specialisation

As stated before there are two levels at which a team can choose to specialise itself. That can be on the individual level or the team level. Each level has distinct reasoning and consequences.

First one needs to understand why specialisation would need to take place. Again, for each level there is a distinct reason. An individual scientist is an organ working within a specific scientific field. [SFE15] [dRP18] Each field can be divided into a finite amount of sub-fields. An individual can focus on one sub field and become a master in this

field. [Ver16] Which results in a deep understanding of this sub field. This introduces the option of labour division on a team level. Where each scientist is an master at a sub field, combining these scientist will get an in-depth comprehension of the entire field. Therefore specialisation at the individual level and division of labour on a team level makes teams more efficient. This efficiency benefits the creative process.

Furthermore, a specialised team tries to get a focused and deep understanding of one specific discipline. For example a peanut allergy specialist is going to collaborate with a walnut allergy specialist, both are nut allergy scientists, but in a different niche. Both are specialists and the team is specialised in one field(nut allergies). We say that a team is specialised when there is focus on some knowledge categories. The same applies for the individual. Moreover the focus on one field increases productivity and visibility, more than a diversified team would. [LBS16]

Secondly one wishes to know the effects of specialisation at each level. For the individual scientist specialising in a niche science has some benefits. By specialising one's world becomes smaller as one becomes part of an island rather than the entire domain. [LR08] By doing this the chance that one's work gets noticed is more likely. [dRP18] [Lea07] Furthermore, one gets the chance to master this niche field and gain in depth knowledge. [Ver16] The combination of increased visibility with increased knowledge depth results in a higher degree of quality,an increase in research success and higher productivity. [Lea16] [Lea07] [LBS16] [Lea07] Moreover when one has a lower degree of specialisation, he will find hinder in productivity, since mastery has become harder, moreover the field has become broader which hinders visibility, which translates to a negative mutation in salary. [Lea07]

Now one wishes to look at the team level. The first effect of a homogeneous configuration is that of increased productivity. [AGT16] While a collaboration improves productivity most of the time due to the division of labour, in the case of homogeneity this is increased even further. One has several masters working on one product. This also means that discussions, opinions and experience of several masters are converging into a single product. Moreover they speak a common language and are agreed on field convention. This minimalises communication bottlenecks and some aspects of the 'agnecy costs'. As a result, the overall quality of the paper is increased and follows a clear convention. [Lea16] [APV16] [Ver16] This in turn is beneficial for the creative product.

2.1.4 Diversification

Knowledge diversification is when a broad spectrum of knowledge categories is represented. A well-diversified on the individual and/or team level is increasingly recognized as a solution to today's problems. [LWW14] [WTG15b] Today's problems don't listen to the borders scientists have created in the academic world through hyper specialisation into niche disciplines. While such specialisation may indeed lead to far better understanding of a scientific field and increase productivity, it fails to push teams into innovative directions, which are needed to excel at solving complex problems. [LR08] Such diversification can, again, be done at two levels:' the individual and the team level'. First the reasoning for diversification on the individual and team level is discussed. After which the properties, benefits and costs of each level is discussed as well.

When an individual has opted for the 'learn more' choice, rather than specialise, he is called a generalist. [Jon09] The generalist is more able to tackle modern problem statements more effectively. [LR08] [WTG15a] This has to do with the ill defined and cross field nature of modern problem statements. A generalist consequently possesses broad spectrum of knowledge. This comes at a cost of knowledge depth, since the knowledge burden forbids in depth knowledge without specialisation. [Ver16] Recall that diversification on a team wide level cannot be a substitute for individual diversification, which still yields the best results [Ver16]. Thus, a scientist can be tempted to choose to enter new fields, either by choosing another specialisation or just generalise in a more superficial manner in

multiple fields. [Jon09] This leads to personalities. In figure 2.2 the basic personalities and thus diversification options are visualised. We distinguish four personalities. A specialist, specialising with in depth-knowledge into a single discipline; A generalist gathering common knowledge in many fields; A T personality a generalist with less number of fields than normal, but with a specialisation; and finally, the N-personality which is two specialisation fields and if existing common knowledge in the fields adjacent. [ho18]

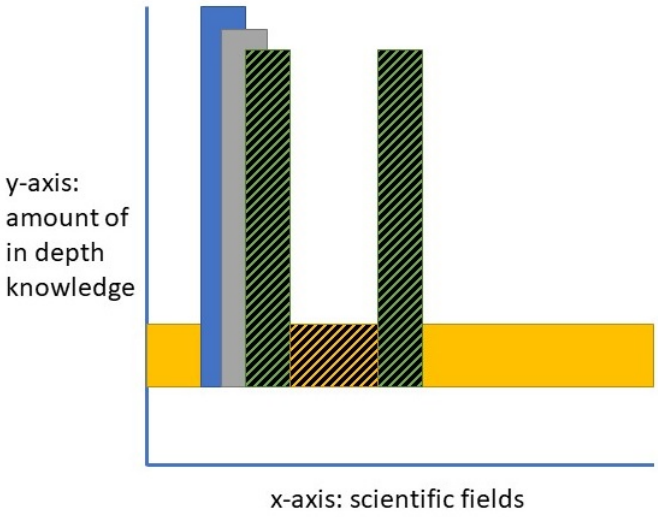


Figure 2.2: The marked area is the N personality, the yellow a generalist, the blue a specialist, a part of the yellow combined with the grey the T personality

These personalities thus effectively rule out the assumption that a scientist must be either a generalist or a specialist(mutual exclusion). Moreover, this illustrates that specialisation and diversification can co-exist.

On a team level one diversifies for more or less the same reason. One wishes to obtain a broad spectrum of domain knowledge in order to increase innovative potential. By introducing dissimilar team members a team can gain a diversified knowledge set in terms of variety. [Ver16] This increases information, which is beneficial for the creative product. On a downside this does slow down productivity. This has to do with an unclear identity, learning cost, increased agency cost and language confusion [LR08]

Now the overall properties, benefits and costs are discussed. The reported fruits plucked from diversification are conflicting. [WTG15b] This has to do with the ambiguousness of the term diversification. Different measures are used, which leads to different results. The paper by Wang, Thijs and Gallnzel [WTG15a] shows that rather than coming up with one arguable measure, one should look at three dimensions of diversification. From this work one dimension is used, namely that of variety. Variety measures how many different niches of science an entity features. One would do well to note that other definitions can be used for diversification such as gender, wealth, age and academic rank, however this paper sticks to knowledge based on categorical variety.

Researches investigating diversification on basis of variety have found that it would yield more citations on the long run and deliver fresh insights. [WTG15b] Whereas the short run shows a negative relation. [LWW14] The increase in citations can be attributed to fact that diversified work leads to an increased change of novelty, and novel work has the change to attract more citations as a reward. Literature offers some explanations as to why citations count is higher on the long run and not the short run. First is the recalcitrant of the scientific establishment towards radicality. This results in slow moving works. Secondly a longer time is needed before follow-up research is conducted. [WVS16] Combine this with a delay of recognition for novel work and the result is that it takes longer

for a novel work to be saturated citation wise. Finally, the citation reward system takes longer to reach its full potential in a 'foreign' field than one's 'home' field. Which further delays full citation saturation. [WTG15a]

2.1.5 Diversification versus Specialisation

One may have noticed that both specialisation and diversification leads to an increase in citation count of the creative works. While this might seem contradicting, it is not. A diversified work is cited more times than the specialised work on the long run. The specialised work is cited more times than the diversified work on the short run. The increase in the citation count is for each variable distinct. Specialisation improves productivity and visibility. As a result the works of a specialised individual gets noticed and attracts citations. A diversified work on the other hand takes longer, for reasons already explained, but it not due to increased visibility. Increased visibility is a trait of a specialised individual. The generalist gets rewarded through the citation reward system, due the novelty of his work. This reward of increased citations outweighs the decrease due to decreased visibility and productivity. In conclusion:

$$Short\ run:\ specialsed > diversified,\ due\ to\ \Uparrow Visbility \rightarrow \Uparrow citations$$

$$Long\ run:\ diversified > specialsed,\ due\ to\ \Uparrow Information \rightarrow \Uparrow Novelty \rightarrow \Uparrow citations$$

2.1.6 Heterogeneous and Homogeneous

Apart from knowledge categories represented in teams through its configuration of individuals, which dictates diversification and specialisation, a team also has multiple similarity configurations, dictating homogeneous and heterogeneous teams. This tells us something about the overlap in expertise of team members, thus in terms of their field of speciality.

An interesting positive relation between the dissimilarity of team members and citation performance on the long run has been found. [WTG15b] [LWW14] Or rather have found a positive relation in combining peculiar knowledge pieces in a creative and meaningful way. [AZM09] Which leads to the claim that diverse knowledge is the prime source for frontier advancing inventions. [Ver16] Another property of a heterogeneous team which strengthens this claim, is that a diverse team members are forced to go through the Storming face of team development and come to an agreement which potentially leads to new innovative approaches, whereas a homogeneous configuration risk going back to convention and so produce more of the same. [NKV17] Furthermore, diversifying team members means bridging otherwise alienated islands of sciences, which leads to better overall results. [LBS16] All of this is of benefit to the finished creative product

There are, naturally, drawbacks as well. First cross field collaboration often means cross field problems. This leads to difficulty in labelling the work and thus creates an ambiguous identity which leads to devaluation. [LBS16] Secondly investing in multiple fields dilutes in depth knowledge and so quality suffers. [Ver16] Finally, there are cognitive and collaboration challenges which are increased in novel teams. [APV16] The best way to characterize this challenge is with the tower of babel. In the early days it is told that all humans were one big group, descendants of Noah. They decided, against their commands, that they were to build a tower reaching in to the sky, so that they could investigate the stars and be the mightiest of all. God would have none of the humans disobedience and made it so that every sub group spoke a different language. As a result every language group broke apart from the main group to form their own society. This effectively illustrates the dangers of a diversified team. Members who speak the same field language and share the same frame of reference are inclined to work together and isolate themselves from the rest of the group. This example is further effective in illustrating language confusion, which exist in team, where team members are of different backgrounds. As a result a diversified team combats communication errors and wasted resources, which have been dedicated to the wrong subjects, due to misunderstandings within the

team.

Finally, recall that diversification on the individual level gains the best results for novelty. The same research also finds that 90% of variance in knowledge is explained by the most diverse team member. [Ver16] This implies that most teams have a generalist, while the others are specialist. While it is unclear what the exact effect of this particular configuration is, it would imply that this would be a success story, why else would so many team incorporate a generalist with many specialists? A possible explanation is given by E. leahey [LR08], where he writes about the fragmentation in sociology, due to hyper-specialisation. He states that specialisation may serve as a *"springboard for innovative work"*, when multiple specialist complement each other or if a generalist joins up as new topics are investigated. [LR08]

A homogeneous team, on the other hand, is the opposite of the heterogeneous team, here the members are similar. They speak a common language and know one field very well, with all its conventions and demands. Moreover they are masters in that field. As a result they are more productive and more known due to their clear identity and field boundaries. While this is beneficial for the quality and therefore usefulness of the final work, it lacks innovative potential. [Lea16] They risk to fall back to more of the same, due to the absence of question and discussion sparked by new perspectives and the overall agreement on field related approaches. Moreover they posses but a small spectrum of knowledge. It is a very powerful spectrum with in depth knowledge. It is however not more information, which is not beneficial for novelty.

2.2 Problem Statement

To summarise the theory thus far. Due to the ever-complex academic world in combination with the knowledge burden, scientists are forced to specialise. The specialising of academics has contributed to the creation of disciplines and sub-disciplines. These disciplines, while some lie in overlap, are inclined not to interact with each other, creating islands of science. Collaboration in these islands is common, while cross collaboration is rare. However, cross collaboration is on the rise, since it is increasingly valued for its novel and useful character. Collaboration in general is needed due to the vast amount of knowledge present, which a single person can't comprehend. Apart from this necessity collaboration yields other advantages as well, such as: 'cost sharing; increased mobility due to modern technology; learning opportunity; possible synergy; increased problem-solving capabilities and higher publication productivity', this also comes at the cost of: 'effectiveness; higher administrative cost; possible agency costs; differences in risk/benefit trade off and communication costs'.

Collaborations are increasing as has been the demand for generalist individuals in science. These generalists provide fresh insights, innovative ideas and a unique way of thinking. The most important reason for the demand however, is their capability to speak multiple 'science languages', such that the generalist can form a bridge between alienated disciplines.
Moreover collaborations can be marked as specialised and generalised by looking at the distribution of its members. The degree of specialisation and generalisation can be computed by measuring diversification on basis of variety.

A diversified team has higher citation count on the long run, than its specialised counterpart. However, on the short run the specialised work is cited more times. A possible reason is delayed recognition or a slow-moving foreign field. The reason for more citations is that a diversified team work is more novel and thus cited more. Citing is a primary part of the research reward system. Thus, a diversified team seem to produce higher valued work. Other research on diversity has shown that knowledge variety has an increasing relationship with citation count. Increased knowledge variety implies increased information amount, which is positively associated with an increase

in novel potential.

The main goal of a research is to expand the knowledge frontier. The knowledge frontier expands when new knowledge is added to the scientific memory. Such a breakthrough can be partly credited to novel work. However novel work does not automatically generate new knowledge, since it needs to be useful and even then, a push forward is not guaranteed. Both usefulness and novelty are important aspects of the creative product. Creativity is puzzling with pre-existent knowledge into a unique whole or the decision to try a radical new approach, which if successfully applied can lead to novel and useful work.

The following research question is introduced based on the theoretical background.

> What kind of team configuration in terms of ability, diversification/specialisation and knowledge overlap benefits creativity the most?

2.3 Team Configuration

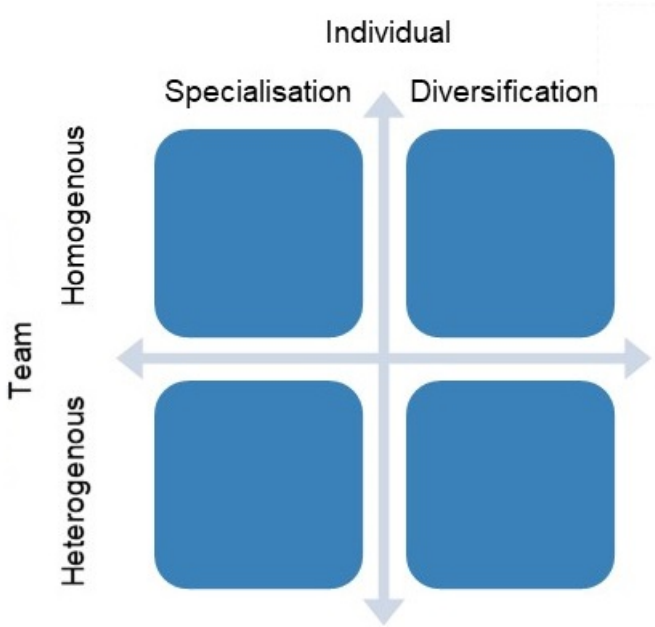


Figure 2.3: The theoretical framework dimension in a matrix format

From the previous theoretical section it was evident that one can focus or expand in terms on knowledge and similar concepts on two levels. That of the individual level and that of the team level. On the individual level one could specialise or diversify, on a team level one similarity and disimiliairty. This is the framework of team configuration, which is visible in figure 2.3. This research takes the team configuration effects under the loop, by conceptually defining three dimensions of the team configuration:' ability, knowledge diversification and knowledge overlap'.

2.3.1 Member abilities

A team effort consists out of multiple individual scientist, with the minimal number of two. Each member has a skill set and certain traits. Where traits are passive, thus something one is born with or gains through certain experiences, and skills are dynamic, which are certain know-hows, approaches etc. which one can learn and master,

but most importantly forget again. [Dic17a] [Dic17b] The word ability is used to describe skills and traits under one term. In this research abilities define the quality of delivered work, its productivity and how creativity is applied. All three have a great impact on research performance. [NKV17] [AZM09] [WJU07] Adjacent works have found that previous performance may define future performance. This has to do with experience received from prior research and path dependency. These experiences hone, define or add new abilities. [NKV17] [AZM09] Furthermore they also find that impact of prior work is often used as promotion and funding purposes. [WJU07]

Combining these findings, this research defines the terms individual mean ability and individual ability variance. The mean ability refers the mean of prior experiences individual scientists have received in their career within a collaboration effort. It states how able a team is to promote their research, receive funding, create quality work and produce research at a fast pace. Furthermore, mean ability would also state how able a team is to solve uncertainties and unforeseen consequences by pulling from past experiences.

The distribution of abilities of individual scientists within a team collaboration shows how homogeneous the team members are, see section “Heterogeneous and Homogeneous”. A large variance within the distribution would suggest that there are some able scientists, but also relatively incapable scientists as well. These incapable scientists would keep the capable ones down. Since extra costs must be made to accommodate learning costs, quality control and extra communications.

Based on previous sections the hypothesis sounds the following:

H_{a1}: ‘A capable team will be more creative
H_{a2}: ‘A homogeneous team in will be more creative’

2.3.2 Specialist or Generalist

Besides traits and abilities, a team’s individual also has a certain knowledge type or personality. Recall the generalists, specialists, T and N personalities which were discussed in previous sections. However, due to measurement limitations the T and N personality cannot be measured effectively. Each of the personalities however, has a distinct effect on team performance. [Ver16] To measure the team’s performance on the subject of knowledge, the variable individual diversification mean and individual diversification variance are introduced. Both measures are related to the variety dimension of diversification introduced by Wang, Thijs and Gallnzel [WTG15b]. Recall that a diversified team would bring more innovation and processing capabilities. Furthermore, diversification is the groundwork for novel perspectives. [LWW14] Which is determined by the overall represented knowledge categories. [dRP18] This means that a high individual diversification mean would suggest that many knowledge categories are present within a collaboration. This would increase knowledge pieces available which can be used in innovative and creative ways. [LWW14] Whereas a low mean would suggest that the team collaboration effort focuses on a handful of categories, thus giving reason to speak about a specialisation. Recall that specialisation has its own benefits such as increase productivity, in depth understanding and increased visibility.

The variance in the distribution of the knowledge categories represented gives us an idea how the team is configured. A high variance would suggest that the individuals within a team would represent different knowledge categories. Whereas a low deviation would suggest of a team, which has the same knowledge accumulation. From previous sections and variables introduced here the following hypothesis is formed.

H_{b1}:’A higher diversification mean will be more creative’
H_{b2}:’A focused, homogeneous, team will be more creative’

2.3.3 Knowledge Overlap

A team consists out of individuals as we stated before, however perhaps a better definition would be that a team consist out of all unique author pairs one can create with the individuals. In other words a team consists out of one or more sub teams in form of all possible author pairs. This opens a new door to evaluate the worth of a collaboration. One now can measure similarity and cognitive distance. The paper:' at the origins of science' [APV16], proposes similarity or distance in knowledge as a measure of cognitive distance. This can be used to measure the knowledge sharing that takes place in collaboration. Where a similar author pair which has a low cognitive distance, would result in that it is easier for a scientist to learn from the other author, since they speak a common language, or in other words they share the same frame of reference. [APV16] [WTG15b] On the flip side of the coin is high cognitive distance or dissimilarity between authors. With the high distance the chance that the scientist actually learns something useful is increased, because of the more novel knowledge one is able to learn, or as the original paper puts it: 'it opens new horizons'. This paper also finds an inverted U relation to cognitive distance. [APV16]

Another property of collaboration is the knowledge overlaps measure. The knowledge overlap would capture expertise within a collaboration. By combing the distribution of the overlap with the mean overlap of pairs within a team, one can find out how specialised a team is on one subject. When they all share the same knowledge the deviation in distribution would be low and the mean high, whereas when some share the same knowledge the mean would be high, but the deviation would be high as well. Furthermore, one can infer the tactic used in team configuration. Two tactics can become evident from a similarity measure of knowledge. One tactic is the complementary tactic. This is when authors try to fill in for each other to create one big expert entity on multiple fields. This tactic offers synergy possibilities. The other tactic is that of reinforcement. A reinforcement tactic is when multiple specialists in the same field harness themselves into one field. This tactic offers powerful, precise problem solving at a fast pace, one can compare it to a surgeon's scalpel. [dRP18] [WVS16]

Based on previous section such as diversification and specialisation and the measure introduced in this section the following hypothesis is formed.

> H_{c1}:'A mean of pairwise similarity on the basis on knowledge categories will have a positive impact on creativity.'
> H_{c2}:'The distribution of pairwise similarity on the basis on knowledge categories, will have a negative impact on creativity.'

In the next two figures one can find the hypotheses in form of a model and all conceptual variables with their relation to the two levels of specialisation and diversification. Do note here that the individual ability mean is the odd one out. This is not directly related to knowledge categories. In theory one would want a high ability mean in all cases. Therefore the variable is marked with a + across every cell.

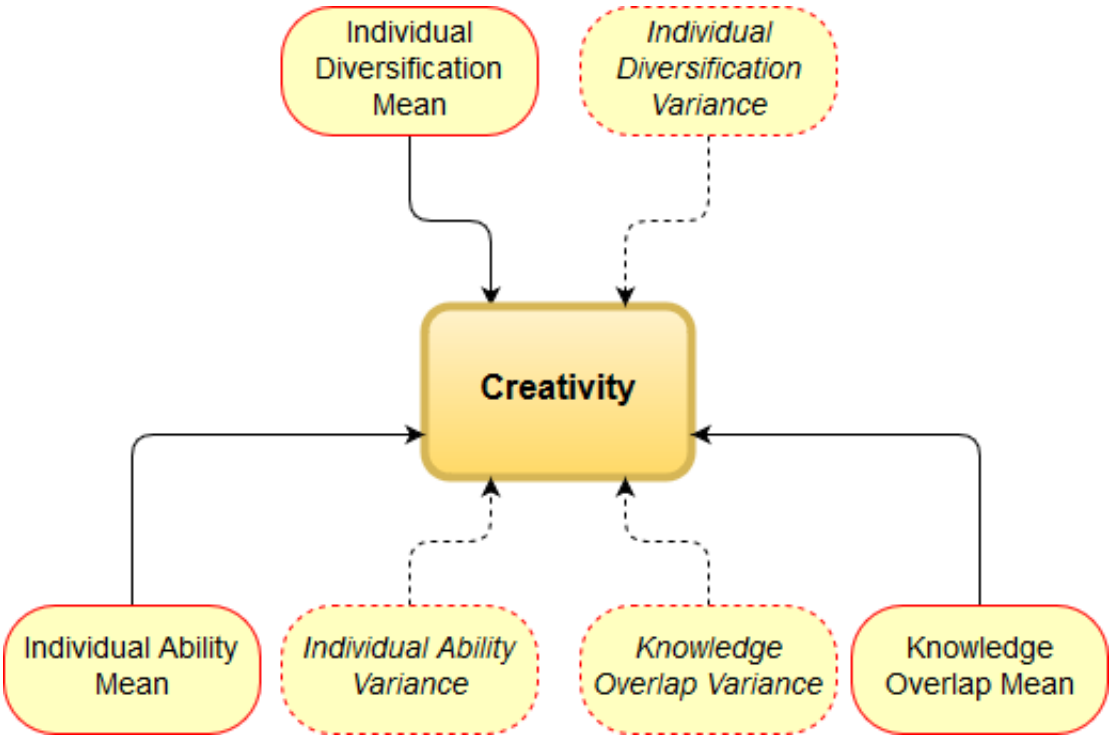
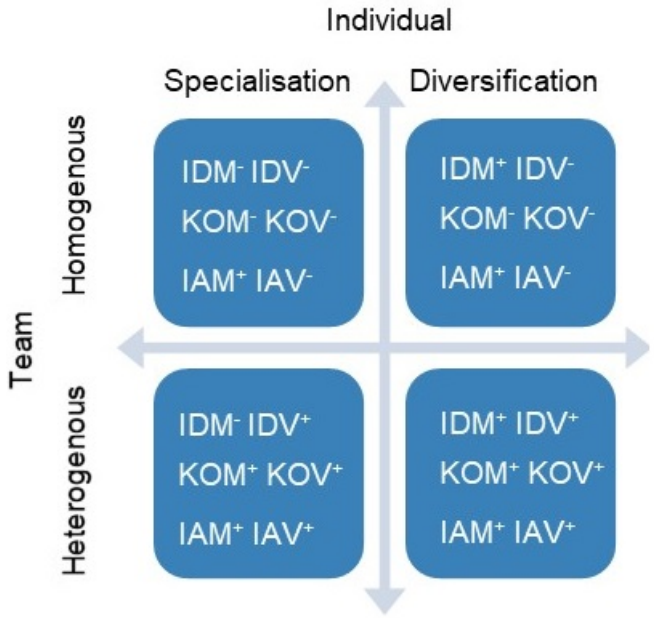


Figure 2.4: The model of the theory -➝= negative relation; →= positive relation



IDM	Individual Diversification Mean
IDV	Individual Diversification Variance
KOM	Knowledge Overlap Mean
KOV	Knowledge Overlap Variance
IAM	Individual Ability Mean
IAV	Individual Ability Variance

Figure 2.5: The theoretical framework dimension in a matrix format, filled with the conceptual variables. A – indicates a low amount, whereas the + indicates a high amount

Chapter 3

Research Design

To investigate the effects of team ability, knowledge diversity and overlap on creative products we focus on the biomedical discipline. The biomedical field is characterised as a fast moving discipline. [SFE15] This means that citation count should be saturated fairly quickly. Consequently we can take a more recent time window. Moreover the biomedical field posses great on-line data-banks, with free access. These data-banks also posses some unambiguous knowledge categories which is a requirement for this research. From these databases the needed data is retrieved.

3.1 Data

Recall that for diversity measures this paper uses is the variety measure for knowledge. [WTG15b] The variety measure tells us of how many different knowledge categories are present. In other words variety is the COUNT function of knowledge types. While the function COUNT is simple indeed, a formal description of knowledge types is not. The natural language is filled with synonyms, homonyms and more importantly the meaning of some words are not agreed upon. Due to the ambiguity of the natural language one cannot simply take the keywords of a paper for instance as valid knowledge types. Therefore one needs to look to a valid syntax of possible knowledge types wich don‘t suffer of ambiguity. Thus, one needs to apply a structured and unambiguous categorical function on works produced by scientists. The MEDLINE sub database, which consist out of biomedical works, of PubMed offers this property in the form of the so-called MeSH terms. MEDLINE has applied a categorical(MeSH) function to each journal paper entry. A MeSH terms stands for *‘Medical Subject Heading‘* and has the following properties:

- Hierarchically ordered
- Unambiguous
- Unique
- Has incorporated synonyms in to one category (part of unambiguous)

A part of the MeSH term database is illustrated in figure 3.1.

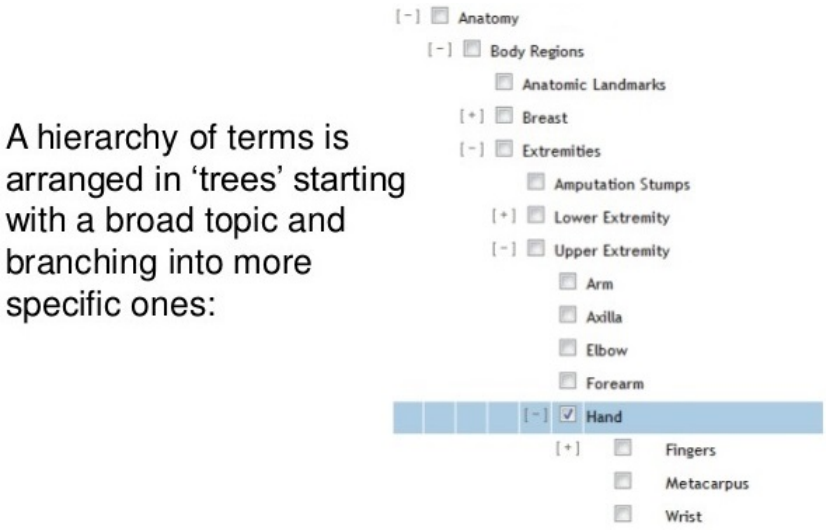


Figure 3.1: A part of the MeSH hierarchical database [Inf14]

Besides MeSH and other information, the PubMed database, subsection Medline, also offers unique identifiers for each paper. This unique identifier is called *'PubMed ID'* and is abbreviated with *'pmid'*. This unique identifier is critical in building the dataset. This gives the opportunity to investigate each work and retrieve all co-authors for each unique paper. From the main PubMed site one can query the following:

$$< year > [pdat]\ AND\ Medline[sb]\ AND\ Journal\ Article[ptyp].$$

This query is used to retrieve the needed database in XML format. The resulting database of $< year >$ will be referred to as <year>XML–DB. In the query the *pdat* refers to the year a paper was published, the *sb* refers to the sub database one wishes to access and finally *ptype* refers to the type of work one is interested in, in our case one is interested in works which were published in journals. The foremost reason for journals is that it is easier to retrieve forward citations and citations of previous work. Our date focus is on the years 2004 – 2007. The year 2007 is our target year in which the creative works are measured. The prior years 2004 – 2006 are used to construct various variables. These variables are named later in this section.

For each year the XML–DB is retrieved and converted to a structured plain text format, namely tsv which is a tab separated csv file. These files will contain two attributes, namely the $pmid$ and the $mesh\ heading\ list$ all else is filtered out. Consequently the following dataset is formed $< pmid > -- < year > -- < MeSH\ terms >$. This dataset is still incomplete as the authors still need to be found and the citation count for each $pmid$ needs to be retrieved.

The retrieval of authors for each $pmid$ poses a problem. While the PubMed database contains author names for each paper, one cannot create a proper author ID. There are three reasons why this would prove a tedious task. First a name is not unique and multiple authors may be named the same. Secondly PubMed has no consistency in naming, for example for one paper the author might be registered with his entire Christian name, while in the other paper only an initial is used. One cannot distinguish this and therefore cannot conduct this research properly. Finally, a person might change his name or increase in academic rank, such that his prefix changes. Therefore, we call upon another database, namely the *'Author-ity* 2009 - *PubMed author name disambiguated dataset'* by V. Torvik and N. Smalheiser [TS18]. This dataset contains a unique author ID, abbreviated with aid, combined with all the pmids an author has worked on, along with other information such as the top 25 mesh terms. However, there is still a problem, a top 25 mesh heading list is not adequate for this research. The reason

for this is when a specialist decides to diversify, then the category in which he has diversified would not show, since it would not be in the top 25. This brings us to the reason why multiple years of XML-DBs were retrieved from PubMed. This so that we can build our own $aid--mes\ list$ database. First we aggregate all aid with their respective $pmid$s in our previousky build dataset. Consequently we now poses the following data structure $<pmid>--<aid>--<year>--<MeSh\ terms>$. From this dataset the years $2004--2006$ are used for the measure of the independent variables. The year 2007 for the independent and control variables. The dataset is still not finished, the citation count for each $pmid$ still need to be retrieved. The citation count is retrieved from the Web of Science(WoS) for each $pmid$ for each $year$. Were 2007 is used to measure creativity per $pmid$ and $2004-2006$ is used for independent variables. Now the dataset is complete, our master dataset.

Now the dataset is filtered, only papers that have proper citation data are chosen and have at least one possible author pair. When however, for each variable a histogram is calculated with 100 bins, an undesired skewness is discovered. To reduce skewness of the database a natural log transformation is applied to the master dataset. However, the dataset contains many zero values, since $\ln(0)$ is undefined the following transformation is used $\ln(x+1)$, which is abbreviated with $log1p$. The new found master table log1p is described in figure 3.2.

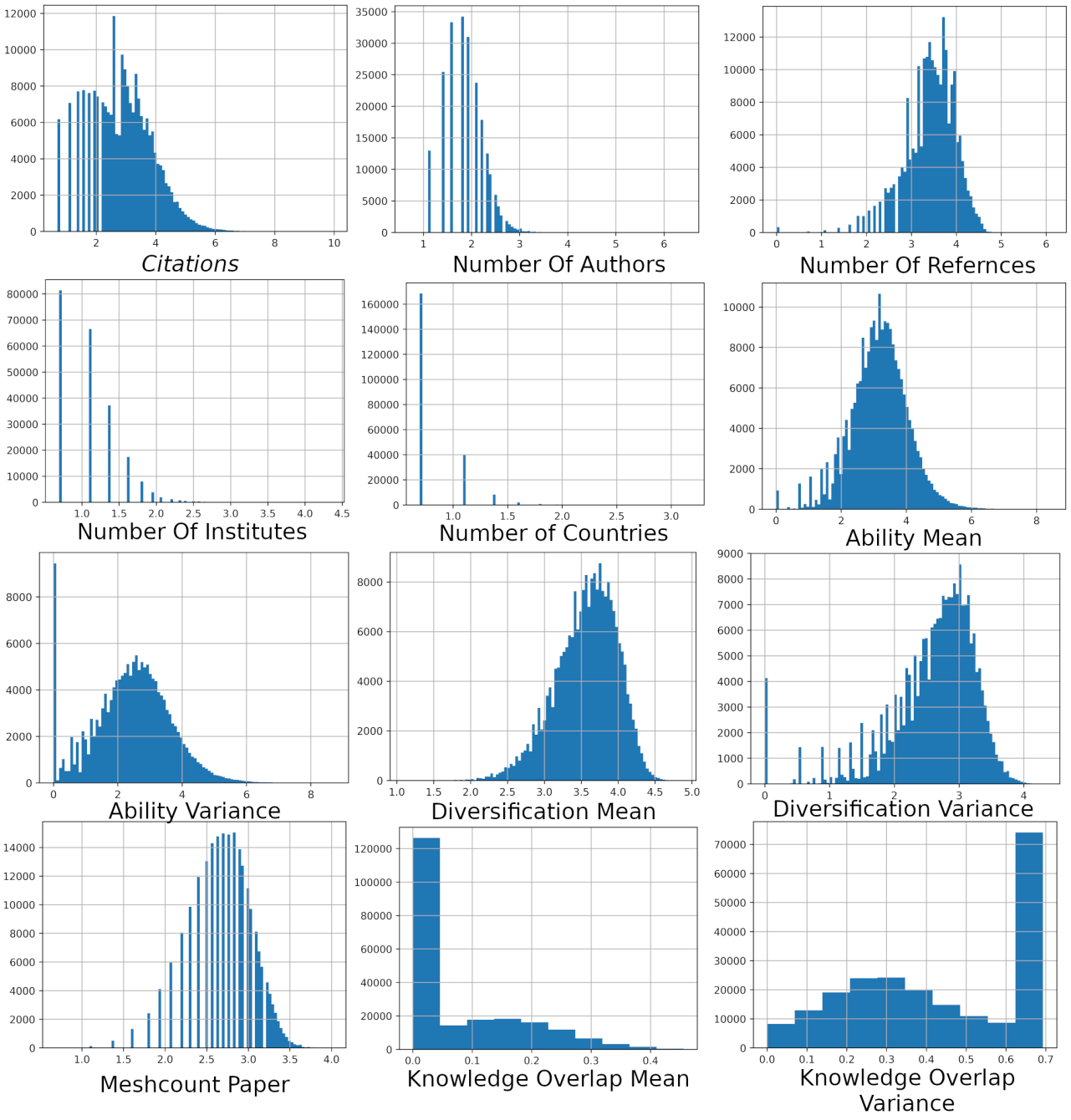


Figure 3.2: Histograms of the independent variables for the log1p master dataset

3.2 Variables

3.2.1 Depended

The creativity impact must be measured to reject or confirm hypothesises. However, the measure of creativity is not that black and white. Creativity is evaluated in the common world by experts and the masses. The same can be applied to science where the scientific community can be seen as the masses through the citation reward system. Therefore, only journal articles are investigated, since citation count can be retrieved. The scientific community evaluates, or rewards published work through citation, since they found it useful and/or novel enough to cite the respective paper. Therefore, the citation count is introduced as the measure for impact of creativity, which is abbreviated with cit.

Creative research is not all roses either, on the contrary it is not for the faint hearted. It is increasingly characterized as high-risk research. [WVS16] [AZM09] This is one of the reasons why grants do not commonly get distributed to radical teams. [NKV17] However, high risk comes with high reward. There is a real chance to advance the knowledge frontier and have a higher citation count on the long run. [WVS16] However due to the high risk, high gain nature of the novel research, citation count shows high variance. Novel work is therefore either highly cited or almost never cited, on the long run. [WTG15b] [WVS16] Literature offers some explanations as to why citations count is higher on the long run and not the short run. First is the most recalcitrant of the scientific establishment towards radicality. Secondly a longer time is needed before follow-up research is conducted. [WVS16] Furthermore, there is a delay of recognition for novel work. Lastly, the citation reward system takes longer to reach its full potential in a 'foreign' field than one's 'home' field. [WTG15a] Recall that another property of creativity is usefulness. When a work is useful, it is used and therefore receives citations. Therefore when a work is useful and novel then the work should be highly cited, since unuseful novel work would lowly cited.

3.2.2 Independent

3.2.2.1 Ability and Skill

Recall from the theoretical framework that ability and skill are often associated with quality, production and performance, among others. Other research suggested that quality and productivity can be measured with citation count. [WJU07] So would this imply that ability can be measured with previous performance? Another prominent research finds that the average received citations is a valid measure to define relative team impact, which correlates with research impact. Also $x \rightarrow y \rightarrow z$ therefore $x \rightarrow z$ under Hypothetical Syllogism rule. So one can say that average previous citations received correlates to research impact. In conclusion quality and productivity can be measured with citation count and average previous recieved citations can define forward citations recieved. Therefore we define ability in this research as the average previously recieved citations. In other words the previous research impact gives us insight into the ability of a scientist. For each scientist the following function is applied between the years 2004 – 2006:

```
 for each aid{
_____total citation = 0
_____works = aid.get(pmid list)
_____ for each pmid in works{
_____ _____total citation = total citation + pmid.get(citation count)
_____}
_____previous citation = total citation / COUNT(works)
}
```

The previous citation which was calculated by this function is then aggregated for each paper and divided by the number of co-authors. This ability mean gives us insight into the abilities of the formed collaboration. A high ability mean would indicate that the author has experience in writing successful papers and is recognized by others. The standard deviation of this measure is called the Individual ability variance. The ability variance gives insight on how experienced the team is. While a large standard deviation would probably be an excellent learning opportunity for those team members who are scrapping the barrel. The question is if this is wanted. On the one hand, this could be hidden talent with fresh ideas and insights, on the other hand, these team members could be a dead weight, since they spend more time making errors while learning, then they are contributing. This research hypotheses that the bad outweighs the good and a low ability variance would probably have the best creative impact.

3.2.2.2 Generalist or Specialist

In the theoretical framework, it was evident that a variety measure for diversification would be used. Variety answers the question of how many types of knowledge one has within a team. In other words, solve the formula COUNT(knowledge categories) or COUNT(MeSH for each scientist). From this the measures individual diversification mean, which is the average of COUNT(MeSH for each scientist), and individual diversification variance, which gives us an idea of the distribution, is introduced. Recall that a team must at least have one author pair. Thus no problems with the standard deviation should occur.

The mean meshcount of the co-authors gives an idea of the overall knowledge that the team's contains. The higher the knowledge the more puzzle pieces the team can utilize creativel. The variance gives insight on how diversified the knowledge is, when the standard deviation is low, then the authors are of the same nature, reducing possible creativity, since no differentiation would lead to no 'new' insights for the individuals in the team, but has higher productivity and less chance of language confusion. Whereas the high variance would have opposite repercussions.

3.2.2.3 Knowledge Overlap

Recall that knowledge overlap is a similarity measure of knowledge categories shared by an author pair within a collaboration. An often-used similarity measure is the cosine measure. [LWW14] This measure is used here as well. For each possible unique author pair of a collaboration the cosine measure is applied. This measures the similarity of the MeSH terms yielded by each respective author. From these results the mean and standard deviation can be calculated. These are respectively called, knowledge overlap mean and knowledge overlap variance. The measures give us insight in to expertise and cognitive distance. For both mean and variance the dataset is filtered further by introducing an extra requirement, which is that each team must have at least three authors, or in other words two possible author pairs.

The knowledge overlap mean gives an indicator to the overlap in knowledge in the team. A well-diversified team would see, some overlap, since proper communication should be possible, however to much overlap would suggest that every co-author would yield the same knowledge and is thus more or less redundant in terms of categorical variety. The knowledge overlap variance gives a general idea of the distribution. Low value means a homogeneous team with low cognitive distance and low expertise, whereas a high value is a heterogeneous team with high cognitive distance and high levels of expertise.

3.2.3 Control

To measure the effect on creativity one must correct for other factors which are correlated with both the dependent and focal independent variables. Thus, from WoS we also retrieve the number of authors for each paper, the disciplines(fields) the paper covers, the number of different countries, the number of different institutes and the number of references. The results are fixed for team size, number of references and fields since Y. Lee, J. Walsh

and J. Wang [LWW14] found that team size has a positive impact on citations, which is partly decided by field and reference variety.

The use of the number of countries and number of institutes as dummy variables is due the fact that these are common measure for calculating diversification on a third level. [JWU08] This level is not interesting for us, but does have effects on citations, thus we control for this effect.
Furthermore, we also control for the novelty and diversification(or specialisation) of the problem statement, by using the knowledge variety measure. This is calulated for each work in the year of 2007 and it called $meshcount\ paper$.

Attribute Name	**Description**
Independed	
mean_prev_cit_authors	The average citations of all co-authors per paper in 2007 between 2004–2006
std_prev_cit_authors	The citations standard deviation of all co-authors per paper in 2007 between 2004–2006
mean_cosine	The average cosine between co-author pairs per paper in 2007
std_cosine	The standard deviation between co-author pairs per paper in 2007
mean_meshcount_authors	The average amount of categories authors of a paper in 2007 have
std_meshcount_authors	The standard deviation of categories of co-authors per paper in 2007
meshcount_paper	The unique count of categories for each paper in 2007
Depended	
cit	The citation count for each paper in 2007
Control	
n_authors	The number of co-authors(team size) per paper in 2007
n_references	The number of references for each paper in 2007
n_countries	The number of countries represented by the co-authors for each paper in 2007
n_institutes	The number of institutes represented by the co-authors per paper in 2007
fields	see table A.2

Table 3.1: Variables in the master dataset explained

3.3 Methodology

In this section the method to retrieve results relevant to the problem statement is described for the sake of reproducibility. This research hopes to investigate six effects. First the effect of only the control variables is investigated, this gives us a frame of reference to interpret the results. Second the effect of ability, individual diversification and knowledge overlap is investigated in the following combinations:

- Ability
- individual diversification
- individual diversification and knowledge overlap
- Ability and individual diversification
- Ability and individual diversification, combined with knowledge overlap

These effects are investigated by using statistical descriptors and regression. Regression was chosen since all variables are continuous, with the exception of the scientific fields. Scientific fields are a control variable which has been converted to an discrete matrix, which can thus be used in continuous inference. Furthermore the meshcount paper has strong correlations with meshcount of authors, thus as a robustness test, all regressions are done a second time without the meshcount paper variable. Results are said to be significant when the t-value is high enough that the ρ value is lower than 0.1. The overall regression is said to be significant if the added independent variables together with control variables are collectively significant, for this the F-statistic is used.

After procuring the dataset, which is described in the section data 3.1, the first thing that needs to be calculated is the correlation matrices, to get an idea of the repercussions of the independent variables. Three correlation measures are used, namely: Spearman, Kendall and Pearson, where Spearman and Kendall or non-parametric measures. After the calculation of the correlation heatmaps the descriptive statistics are calculated for the dataset.

From the information given by the statistical description in combination with the Occam's razor, it has been decided that a linear regression would be most fitting. Occam's razor states that the most simple explanation is prefered over the more complex. In other words the model with the minimal description length, or the minimal bytes, describing the same problem is chosen. Three types of linear regression are taken into account, these are:

- Ordinary least squares (OLS)
- LASSO regression
- Ridge regression

The OLS regressor calculates relations based on the sum of difference from observed and predicted values for the dependent variable. The LASSO regressor extends this behaviour by introducing a cost function of type $l1$. This cost function is called least absolute shrinkage and selection operator and regularises the learning function with a certain strength called $alpha$. The ridge regressor also extends the OLS behaviour with a cost function. However, the ridge cost function regularises towards the $l2$ norm, rather than $l1$. [PVG+11] Each of these regressors is run on the master log1p dataset. After which the most promising one is used for further calculations.

With the most promising regression learner the needed results are computed. To do this the master dataset is divided into pieces, each piece corresponds to a effect one wishes to investigate, each simply named after their conceptual variable name. On each dataset piece regression is applied. From this the p-value, t-value, coefficient, intercept, R-squared adjusted for degrees of freedom and the F-statistic is retrieved. Where possible visualisation is applied. To do this the python scripting language [Ros95] is abused with certain packages. The following python packages are used:

- pandas: this is used for loading the database into memory, data mutation and basic statistics [McK11] [McK10]
- NumPy: this is used for array representation and mathematical functions such as log1p [TE06]
- itertools: for more memory friendly iteration through data structures
- matplotlib: a visualisation library [Hun07]
- seaborn: an interface with visual templates for matplotlib [WBO+17]
- sklearn: machine learning library, used for regression [PVG+11]
- statmodels: statistical library, used for regression [SP10]

The reason that two separate regressor-libraries are used is quite unfortunate. However, before we go into details as why this is, please note that both regressor libraries have been tested and produce the same model on the same dataset, with mere insignificant differences. Consequently the resulting models are equivalent and can be used interchangeably. The reason that sklearn is used is that this library is much easier visualised, supports better data pre-processing and has an inbuilt cross validation function. However, sklearn doesn't automatically computes the needed statistics such as the p-value or the t-static. For the statistical descriptors the statmodels library is used.

For each effect of the six effects that we investigate the same steps are performed.

1. Load dataset into system memory with pandas
2. Validate data for completeness
3. Scale data with sklearn
4. Build model with statsmodels and sklearn on scaled data
5. Export statistical description of the model with statsmodels
6. Run tenfold cross validation with sklearn
7. Export predictive power found with cross validation in 2D space with matplotlib

Step three is standard scaling. This step transforms the data, such that the mean is removed(set to zero) and is scaled to unit variance. The is done for each attribute separately. This makes each feature look like a normal distributed dataset. This is needed since some regressors assume a normal distribution. If a feature would have variance bigger as other attributes in the dataset, then that feature would dominate the dataset. This would lead to biased results and makes the learner in question unable to learn from other attributes.

Chapter 4

Results

In this section the results are presented. This is done in two fold, first the dataset is described, secondly the machine learning regression results are presented. Recall that the most promising linear regressor is chosen to investigate the effects conceptually introduced in the theoretical framework.

4.1 Descriptive

In the table below basic descriptive statistics can be found. Note that the lower count in the knowledge overlap has to do with the requirement that knowledge overlap needs at least three authors or two possible author pairs. This necessary for the measurement of knowledge overlap. On a different note in the heathmap below, the correlation matrix is visible for the Spearman measure. In the appendix other measures such a Pearson and Kendall can be found, see figure A.1 for Kendall and figure A.3 for Pearson. Note from the heathmap that meshcount paper and meshcount authors are strongly correlated.

	cit	n_authors	n_refs	n_institutes	n_countries	mean_prev_cit_authors	std_prev_cit_authors	mean_meshcount_authors	std_meshcount_authors	meshcount_paper	mean_cosin	std_cosin
count	220051.0	220051.0	220051.0	220051.0	220051.0	220051.0	220051.0	220051.0	220051.0	220051.0	162056.0	162056.0
mean	2905	1895	3437	1109	0.81	3162	2434	3566	2606	2.74	436	75
std	1041	414	551	407	234	879	1195	414	686	373	0.22	102
min	693	693	0.0	693	693	0.0	0.0	1099	0.0	693	0.0	0.0
25%	2197	1609	3135	693	693	2639	1.7	3308	2313	2485	249	0.0
50%	2.89	1946	3497	1099	693	3179	2495	3611	2748	2773	417	0.0
75%	3584	2197	3807	1386	693	3714	3214	3871	3068	2996	693	147
max	9967	6435 at least one author pair rule and no new authors for cosine, since this has undefined behaviour	6.14	4344	3178	8471	8732	4866	4339	4007	693	456

at least one author pair rule and no new authors for cosine, since this has undefined behaviour

Table 4.1: Description of the master log1p table

Figure 4.1: Correlation heatmap based on the non-parametric Spearman measure for log1p dataset

4.2 Predictive

4.2.1 Regressor

In table 4.2 one can see the findings of this research. Recall that Ridge, LASSO and OLS were tried on each predictor isolated. If the results would improve with Ridge or LASSO those would be used. If not than OLS, under the concept of Occam's razor. Occam's razor states that the most simple explanation is prefered over the more complex. In other words, the model with the minimal description length, or the minimal bytes, describing the same problem is chosen. Therefore, OLS is chosen, since Ridge and LASSO both introduce extra complexity, without adding additional explanation.

Attribute	0		Ability		Knowledge				Knowledge & Ability			
	Coeff.+ρ	t	Coeff.+ρ	t	Coeff.+ρ	t	Coeff.+ρ	t	Coeff.+ρ	t	Coeff.+ρ	t
Individual Ability Mean			0.36 ☆☆☆☆	116.91					0.36 ☆☆☆☆	115.45	0.36 ☆☆☆☆	104.80
Individual Ability Variance			-0.05 ☆☆☆☆	-17.15					-0.05 ☆☆☆☆	-14.41	-0.05 ☆☆☆☆	-14.92
Individual Diversification Mean					0.05 ☆☆☆☆	16.53	0.06 ☆☆☆☆	16.59	-0.01 ☆☆☆☆	3.20	-0.01 ☆☆☆☆	-3.01
Individual Diversification Variance					-0.01 ☆☆☆☆	-4.641	-0.01 ☆☆☆☆	-3.23	-0.02 ☆☆☆☆	-9.97	-0.02 ☆☆☆☆	-8.01
Knowledge Overlap Mean							0.00 ☆☆☆☆	1.32			0.01 ☆☆☆☆	1.01
Knowledge Overlap Variance							-0.01 ☆☆☆☆	-2.25			-0.01 ☆☆☆☆	-2.30
Meshcount Paper	0.07 ☆☆☆☆	25.59	0.06 ☆☆☆☆	26.01	0.04 ☆☆☆☆	14.57	0.04 ☆☆☆☆	11.62	0.07 ☆☆☆☆	23.20	0.06 ☆☆☆☆	22.83
Intercept	2.95 ☆☆☆☆	1288.43	2.95 ☆☆☆☆	1366.12	2.95 ☆☆☆☆	1289.78	2.97 ☆☆☆☆	1151.65	2.95 ☆☆☆☆	1366.39	2.96 ☆☆☆☆	1232.31
Number of Authors	0.15 ☆☆☆☆	53.24	0.13 ☆☆☆☆	49.07	0.16 ☆☆☆☆	55.24	0.16 ☆☆☆☆	61.61	0.13 ☆☆☆☆	49.30	0.14 ☆☆☆☆	44.62
Number of References	0.28 ☆☆☆☆	111.49	0.23 ☆☆☆☆	95.46	0.28 ☆☆☆☆	109.78	0.28 ☆☆☆☆	134.34	0.23 ☆☆☆☆	95.81	0.23 ☆☆☆☆	44.62
Number of Institutes	0.05 ☆☆☆☆	16.09	0.04 ☆☆☆☆	12.25	0.05 ☆☆☆☆	15.71	0.05 ☆☆☆☆	17.45	0.03 ☆☆☆☆	12.06	0.03 ☆☆☆☆	10.96
Number of Countries	0.07 ☆☆☆☆	24.19	0.05 ☆☆☆☆	18.80	0.07 ☆☆☆☆	23.84	0.06 ☆☆☆☆	25.84	0.05 ☆☆☆☆	18.70	0.05 ☆☆☆☆	16.96
F-stastistic		247.1		386.6		248.8		247.7		380.0		305.9
Numer of Obeservations		220051		220051		162056		162056		162056		162056
R^2-adjusted		0.22		0.31		0.22		0.22		0.31		0.31

Table 4.2: Main results

☆ =$\rho < 0.1$; ☆☆ =$\rho < 0.05$; ☆☆☆ →$\rho < 0.01$; ☆☆☆☆ →$\rho < 0.001$;

All results are fixed for field variety

In the result table 4.2 four classes of results are displayed, namely *'0, Ability, Knowledge and Knowledge & Ability'*. Recall from the methodology that ability was measured with citations received on prior work and that knowledge was measured with categorical variety in MeSH terms for the individual diversification, combined with the cosine measures for knowledge overlap. First the '0' class, this refers to the features for which one wants to control. The theoretical framework introduced these variables as correlated with citation count. The results show that these variables are indeed significant and positive, which is in line with the literature introduced in the theory section. Furthermore, notice that the unique meshcount on paper level is controlled for as well and has a positive relation. As a robustness test the same table was reproduced without the meshcount on paper level, see the table A.1 in the Appendix A.

4.2.2 Member abilities and skills

The ability of a collaboration is measured with the mean and standard deviation of average of total received citations for prior work between the years of $2004 - -2006$, which is abbreviated with individual ability. The results show a notably large positive relation for the individual ability mean. This relation is significant with a ρ value below 0.001. Which would suggest that a team with previous experience in writing relative successful papers, would receive more citations. This is not strange, recall that other literature has reported this as well. Team members with experience in writing, do not need to repeat beginners mistake, have an established social network and know field conventions. Furthermore they have an improved frame of reference, such that they have an easier time connecting the dots of different knowledge pieces to solve a problem statement. In conclusion a team with more experience under their belt would benefit creativity more.

A negative relation is found for the individual ability variance. This implies that a homogenous distribution would receive more citation than the diversified counterpart in previous citations. This further supports the hypothesis in that a cunning team will receive more citations over a team of multiple beginners and one giant. The beginners would way the giant down. This can be partly explained with that the giant needs to allocate team resources, which could be spent on the research itself, to impose quality assurance, teach the beginners and slip into a perhaps unwanted roll of management. In conclusion a team in which the team members are experienced in the same way, a homogeneous team, would be more beneficial for creativity.

Now one is interested if this relation would change if the meshcount paper is removed. In the table A.1 in the appendix A, one finds that the relation remains positive and unaltered in magnitude. In fact, if we go back to the main results table, one finds that the significance and positive relation for mean previous citations remains robust over all the different classes, the same holds for the standard deviation of individual ability. Al with al there is significant evidence that in the biomedical field in the year of 2007 that scientific teams who have prior experience and thus are more able and skilled in scientific research and have the same level of capability have the best impact on creativity. Which leads to useful novel work, which is more likely to advance the knowledge frontier. Thus the H_0 is rejected and we accept the H_a, thus conforming that homogenous abilities and are of importance if one tries to maximise creativity impact.

4.2.3 Specialist or generalist

From the result table a positive relation is visible for the individual diversification mean. This suggests that the hypothesis would be correct in that a more diversified team members in terms categorical variety would be more creative. Thus, a team with many knowledge pieces on average would excel at creative puzzling due to more available knowledge pieces. Moreover, a negative relation for the diversification variance suggests that team members should be the same. Hinting that a team of generalists, generalised in the same degree would receive a higher citation count. The positive relation for the mean rules out the team of specialists. This is remarkable in

that such a relation cannot be entirely explained with the literature introduced in the theoretical framework.

Indeed, more knowledge pieces in terms of categorical variety improves the accumulated citations. However, it is not the way we expected it. Well other work has established that variance in knowledge is explained for 90% by the most generalised member, implying that others are specialists, we find that a team of generalists would perform better.

The findings however are not robust. When the ability measure is introduced, the individual diversification mean flips and becomes a negative relation. Without ability individual diversification mean has a small but positive correlation. Meaning generalists are in general more able. Without controlling for author ability, as generalists team performs better, but when controlling for ability, a team of specialists seems to perform better. Thus, there is a negative relation for the individual diversification mean, a negative relation for the individual diversification variance and positive relation for the meshcount of the paper. A negative relation for the individual diversification variance suggest that a focus on niche sciences are beneficial for creativity. Furthermore, as an unexpected turn of events, these specialist teams do not need to be diversified, on the contrary the biomedical field seems to benefit from the reinforcement tactic. A team of multiple specialist specialised in the same niche has the best creativity impact. All with all we fail to reject the null-hypothesis for the individual diversification mean, but can reject it for the individual diversification variance.

4.2.4 Knowledge overlap

The knowledge overlap mean tells one if team overlap is either wanted or something to be avoided. A positive relation would suggest a higher amount of overlap would be desired, a negative the opposite. From the result table one can retrieve an indifferent relation for the mean cosine suggesting that overlap doesn't matter for creativity impact. This quite unexpected, but this means that we fail to reject the null hypothesis for the knowledge overlap mean.

From the same table one can find a negative relation for the knowledge overlap variance. This differs from the knowledge overlap mean in that the mean would measure the amount of overlap and the distribution measures the difference in overlap between authors pairwise. A negative relation would then suggest that overlap on different fields would be costly, rather one should have overlap in the same level. This can be explained as that a team needs common ground. Something all can relate to. The relation is negative and significant, we reject the null hypothesis and accept the alternative.

When both knowledge overlap and diversification are considered, the results suggest that a team of generalists, who are diversified in the same knowledge categories and have the same overlap pair wise have the greatest impact on creativity, while working on a problem statement that crosses multiple knowledge categories. When, the ability is introduced, an observation must be made. The knowledge overlap means has become insignificant. This doesn‘t change much for the policy makers, since the mean was first indifferent.

In conclusion a team of specialists, specialised on the focus of knowledge categories and have the same pairwise overlap and have equal previous experience retrieved from prior work, will have the best impact on creativity.

4.2.5 Control Variables

As expected from the literature research the control variables are all relevant and positive. The number of institutes seems to have positive relation and stays significant cross classes. The team size and references also have a significant positive relation cross class. What is interesting is that the team size is less relevant under the introduction of the member ability and skills but increases in prevalence when knowledge variety is added. Recall from the theoretical section that team size was correlated with an attempt to diversify. This shows in the results. Furthermore, notice that all regressions are significant with the increasing F-statistic.

All with all a team of specialists, specialised in the same focus of knowledge categories and have the same pair wise overlap, will have the is the most beneficial for creativity if working on a problem that crosses multiple knowledge categories. One wonders as to why this relation is evident, since theory would point at diversified team not specialised as prime source of novelty and usefulness. One could argue that not enough time has passed, since novelty only performs better in the long run. However, it has been ten years and the biomedicine world move fast, which would suggest that ten years is enough. Another explanation is that 2007 was a notable unlucky year for novel teams. Recall that novel work had greater variance in citation count, due to is high risk nature. So one could speculate that many diversified teams have created novel work, but it was not useful and thus makes up for a low citation count. A far more likely explanation is that the biomedicine world still profits far more of hyper specialisation than it does from diversification or that these fields don't have many diversified teams

Chapter 5

Conclusions

In the biomedical field three effects have been investigated to answer the question, *'Which team configuration in terms of ability, skill, knowledge overlap, specialisation and diversification is most likely to have creativity impact in the biomedical field?'*. Literature finds that creativity is evaluated by experts and the masses. In addition the system of citing and be cited can be seen as a reward system, through which scientist can show their appreciation. Therefore creativity can be measured by citation count. Three conceptual variables were investigated and their effect on creativity. First ability, literature found that ability determines quality, productivity and performance. These are often measured with research impact. Therefore, we argue that ability in a team configuration can be measured by the average and distribution of the research impact of scientist's prior work within the collaboration. We hypothesised that a capable team which is homogeneously distributed in terms of ability would be more creative. Results showed a constant positive relationships across all specifications for the average ability and a negative relation for the standard deviation. Which translates to that a team with with team members who all have had proper prior experiences has the best impact on creativity. Results are significant; thus, we reject the null hypothesis and accept the alternative.

Secondly the effect of generalist or specialist was investigated. This was measured with the diversification measure of variety. Which simply states how many knowledge categories are present within a team. We argued that MeSH terms from the MEDLINE dataset retrieved from PubMed fulfils the criteria for knowledge categories. Therefore, a mean count of all mesh terms accumulated by all team members and its standard deviation would be a proper measurement for individual diversification, which gives us insight in specialisation and diversification of team members. From the literature which found that 90% of the variance in knowledge variety is explained by the most diversified member and that diversified knowledge has a positive relation with creativity impact, the hypothesis was formed that, a generalist managing a team of specialists who each have their own distinct specialisation, would be more creative. Results show two observations. The first observation is in isolation of the other variables, where the individual diversification mean has a positive relation and the individual diversification variance a negative relation. Which suggest that a team of generalist, which have many knowledge categories, thus benefiting from the positive relation of the mean, who have the exact same knowledge, which benefits from the negative relation of the distribution, is more creative. When the ability variable is introduced to knowledge diversification individual diversification mean becomes a negative relation. This effectively means that a specialised team, meaning it focuses on a few mesh terms, specialised on the same categories has the best likelihood to have the best impact on creativity. This result is significant, meaning that we fail to reject the null hypothesises.

In third place is knowledge overlap. Theory suggests that knowledge overlap gives us insight in cognitive distance

and expertise. We argued that the common cosine similarity measure gives us insight into this matter. Theory suggested that the similarity in knowledge stock was a valid measure. Therefore, the cosine measure based on MeSH terms of each possible unique author pair within a collaborative effort is calculated from which the mean and variance is taken. Results showed that the knowledge overlap mean had an indifferent relation and when the ability was introduced became insignificant. The knowledge overlap variance of the cosine was negative no matter the specification. Suggesting that a team configuration should exist out of author pairs who have overlap on the same knowledge categories. We hypothesised a postive relation for knowledge overlap mean and a negative relation for knowledge overlap variance would be more creative. From the results we fail to reject the null hypothesis for the individual diversification mean, but we can accept the alternative hypothesis for the knowledge overlap variance, as we can reject its null hypothesis.

All with all a specialised team of scientist specialised in the same knowledge categories and sharing the same overlap in knowledge categories have the best possible chance to have a meaningful impact on creativity.

This paper contributes on a theoretical perspective by offering empirical results for the ongoing to discussion, if it's better to diversify or to specialise. This research has found that specialists team researching diversified problems, deliver the best creativity impact for the biomedical discipline. In practice this means that managerial activities and policies can use these findings to maximize novelty and usefulness chance, which helps to achieve the scientific goal of advancing the knowledge frontier.

This research has its limitations. First team context was not considered, since a database doesn't hold these kinds of relations. Despite this, other research have proven that context plays a major part in deciding creativity. Therefore some deviations in the dataset weren't controlled for properly. Furthermore, this research is limited to the biomedical field as presented by the Medline database retrieved from PubMed. To add to this is that only authors that were disambiguated in to the disambiguated dataset of 2009 were taken into consideration. This makes it hard to generalise to the entire scientific community. This research also used logarithmic transformation to reduce skewness, however some research has shown that this could bias the data. How it would bias the data in this particular dataset is unknown [CWL+14] Finally, non-linear regressor would have in hindsight have more promising results but would have been harder to interpret. Especially for the knowledge overlap, a linear relation for knowledge overlap is in hindsight extremely unlikely. To illustrate suppose it has a linear relation, than either a team wich is 100% similar or a team which has no form of similarity would have had the best result. If a team would be entirely similar than what would be the point? If one adds a team member who is completely similar no value is added apart for possible labour division, but that is not even a good reason, because if one adds a more or less similar team member one still gets the opportunity for labour division and gains some new insights are a slight deviation in the frame of reference which benefits creativity. Same reasoning applies for a completely dissimilar team. If none of them have something in common, than there is language confusion. Just as the tower of babel, every group, or in this case every team member will go its own way and nothing is delivered as a team result. Therefore some common ground is needed. In conclusion a linear relation for knowledge overlap is unlikely, its more likely it would be a inverted-U relation.

Which leads to possible future works. The goal of this research was previously to investigate categorical balance and disparity alongside the variety. However due to unforeseen difficulties in data preparation, the research was limited to categorical variety. However further research in disparity and balance is needed as it might shed light on to why diversification flips from team of generalist to a team of specialists.
Finally, the same research could be applied to another field, which preferably would have an equivalent to MeSH terms. With that this research could be done again as well, with a better cosine measure and a broader range of years. Furthermore, there is evidence that log transforming data could make data biased [CWL+14], thus when

this type of research is done again, it might be interesting to try the other transform measures proposed in this paper. The foremost future research would be to investigate non-linear relation, especially for knowledge overlap.

[AGT16] Ajay Agrawal, Avi Goldfarb, and Florenta Teodoridis.
Understanding the changing structure of scientific inquiry.
American Economic Journal: Applied Economics, 8(1):100–128, 2016.

[Ama83] Teresa M. Amabile.
The social psychology of creativity a componential conceptualization.
Journal of Personality and Social Psychology, 45(2):357–376, Aug 1983.

[APV16] Charles Ayoubi, Michele Pezzoni, and Fabiana Visentin.
At the origins of learning: Absorbing knowledge flows from within or outside the team?
GREDEG Working Paper No. 2016-08, 2016.

[AZM09] Pierre Azoulay, Joshua S. Graff Zivin, and Gustavo Manso.
Incentives and creativity: Evidence from the academic life sciences.
NATIONAL BUREAU OF ECONOMIC RESEARCH, Oct 2009.

[BC73] Phillip Bosserman and Diana Crane.
Invisible colleges diffusion of knowledge in scientific communities. diana crane.
American Journal of Sociology, 79(1):180–182, 1973.

[CN17] Creativity and Novelty.
Cambridge Academic Content Dictionary .
Cambridge University Press, 2017.

[CWL+14] FENG Changyong, Hongyue WANG, Naiji LU, Tian CHEN, Hua HE, Ying LU, and Xin M TU.
Log-transformation and its implications for data analysis.
Shanghai Archives of Psychiatry, 26(2):105–109, 2014.

[Dic17a] Cambridge English Dictionary.
Skill, 2017.

[Dic17b] Cambridge English Dictionary.
Trait, 2017.

[dRP18] Gaetan de Rassenfosse and Orion Penner.
The returns to scientific specialization.
in proceeding, 2018.

[ftAoS13] American Association for the Advancement of Science.
Science.
Project 2061. AAAS, 2013.

[ho18] https://www.universiteitleiden.nl/medewerkers/tyron offerman.
Tyron Offerman.
LIACS, 2018.

[Hun07] J. D. Hunter.

Matplotlib: A 2d graphics environment.
Computing In Science & Engineering, 9(3):90–95, 2007.

[hw18] https://www.universiteitleiden.nl/medewerkers/jian wang.
Jian Wang.
LIACS, 2018.

[Inf14] S.A. Infoseg.
MeSH Trees.
Infoseg, S.A., Feb 2014.

[Jon09] Benjamin F. Jones.
The burden of knowledge and the "death of the renaissance man": Is innovation getting harder?
The Review of Economic Studies, 76(1):283–317, 2009.

[JWU08] B. F. Jones, S. Wuchty, and B. Uzzi.
Multi-university research teams: Shifting impact, geography, and stratification in science.
Science AAAS, 322(5905):1259–1262, Nov 2008.

[LBS16] Erin Leahey, Christine M. Beckman, and Taryn L. Stanko.
Prominent but less productive: The impact of interdisciplinarity on scientists research.
Administrative Science Quarterly, 62(1):105–139, 2016.

[Lea07] Erin Leahey.
Not by productivity alone: How visibility and specialization contribute to academic earnings.
American Sociological Review, 72(4):533–561, 2007.

[Lea16] Erin Leahey.
From sole investigator to team scientist: Trends in the practice and study of research collaboration.
Annual Review of Sociology, 42(1):81–100, 2016.

[LR08] Erin Leahey and Ryan C. Reikowsky.
Research specialization and collaboration patterns in sociology.
Social Studies of Science, 38(3):425–440, 2008.

[LWW14] You-Na Lee, John P. Walsh, and Jian Wang.
Creativity in scientific teams: Unpacking novelty and impact.
Research Policy, 44(3):684–697, Nov 2014.

[McK10] Wes McKinney.
Data structures for statistical computing in python.
In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010.

[McK11] Wes McKinney.
pandas : powerful Python data analysis toolkit, 2011.

[NKV17] Daniel Neicu, Stijn Kelchtermans, and Reinhilde Veugelers.
Off the beaten path: What drives scientists entry into new fields.
in proceeding, Apr 2017.

[NL99] Sev V. Nagalingam and Grier C.I. Lin.
Latest developments in cim.
Robotics and Computer Integrated Manufacturing, 15:423–430, Jan 1999.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay.

Scikit-learn: Machine learning in Python.
Journal of Machine Learning Research, 12:2825–2830, 2011.

[Ros95] Guido Rossum.
Python reference manual.
Technical report, CWI (Centre for Mathematics and Computer Science), Amsterdam, The Netherlands, The Netherlands, 1995.

[SFE15] Feng Shi, Jacob G. Foster, and James A. Evans.
Weaving the fabric of science: Dynamic network models of sciences unfolding structure.
Social Networks, 43:73–85, 2015.

[SP10] Skipper Seabold and Josef Perktold.
Statsmodels: Econometric and statistical modeling with python.
In *9th Python in Science Conference*, 2010.

[TBM97] D.a. Thurman, D.m. Brann, and C.m. Mitchell.
An architecture to support incremental automation of complex systems.
1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation, 1997.

[TE06] Oliphant Travis E.
A guide to numpy.
Technical report, USA: Trelgol Publishing, 2006.

[TS18] Vetle I. Torvik and Neil R. SMalheiser.
Author-ity 2009 - pubmed author name disambiguated dataset, 2018.

[Ver16] Dennis Verhoeven.
Potluck or chef de cuisine, knowledge diversity and award-winning inventor teams.
in proceeding, Jun 2016.

[WBO+17] Michael Waskom, Olga Botvinnik, Drew O'Kane, Paul Hobson, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian de Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, Chris Fonnesbeck, Antony Lee, and Adel Qalieh.
seaborn: v0.8.1.
Technical report, The seaborn project, Sep 2017.

[WJU07] S. Wuchty, B. F. Jones, and B. Uzzi.
The increasing dominance of teams in production of knowledge.
Science AAAS, 316(5827):1036–1039, 2007.

[WTG15a] Jian Wang, Bart Thijs, and Wolgang Gallnzel.
Interdisciplinarity and impact: Distinct effects of variety, balance and disparity.
PLoS ONE, 10(5), May 2015.

[WTG15b] Jian Wang, Bart Thijs, and Wolfgang Glanzel.
Interdisciplinarity and impact: Distinct effects of variety, balance and disparity.
PLoS ONE, 10(5), 2015.

[WVS16] Jian Wang, Reinhilde Veugelers, and Paula Stephan.
Bias against novelty in science: A cautionary tale for users of bibliometric indicators.
NATIONAL BUREAU OF ECONOMIC RESEARCH, Apr 2016.

APPENDIX

Figure A.1: Correlation heatmap based on the non-parametric kendall measure for log1p dataset

Figure A.2: Correlation heatmap based on the non-parametric spearman measure for log1p dataset

Figure A.3: Correlation heatmap based on the pearson r measure for log1p dataset

Attribute	0		Ability				Knowledge								Knowledge & Ability							
	Coeff.+ρ	t	Coeff.+ρ	t	Coeff.+ρ	t	Coeff.+ρ	t	Coeff.+ρ	t	Coeff.+ρ	t	Coeff.+ρ	t	Coeff.+ρ	t	Coeff.+ρ	t	Coeff.+ρ	t	Coeff.+ρ	t
Mean Prev. Citations			0.33 ☆☆☆☆	140.79	0.36 ☆☆☆☆	117.13									0.36 ☆☆☆☆	116.79	0.36 ☆☆☆☆	114.79	0.36 ☆☆☆☆	104.97	0.36 ☆☆☆☆	104.97
Std. Prev. Citations					-0.05 ☆☆☆☆	-17.06									-0.05 ☆☆☆☆	-17.48	-0.05 ☆☆☆☆	-14.48	-0.05 ☆☆☆☆	-14.92	-0.05 ☆☆☆☆	-14.92
Mean Meshcount Authors							0.07 ☆☆☆☆	28.48	0.07 ☆☆☆☆	28.94					-0.02 ☆☆☆☆	9.61	-0.03 ☆☆☆☆	11.00				
Std. Meshcount Authors									-0.01 ☆☆☆☆	-5.17							-0.02 ☆☆☆☆	-10.56				
Mean Cosine											-0.01 ☆☆☆☆	-3.99	-0.01 ☆☆☆☆	-2.81					0.01 ☆☆☆☆	2.30	0.01 ☆☆☆☆	2.30
Std. Cosine													0.01 ☆☆☆☆	3.13							-0.01 ☆☆☆☆	-14.87
Intercept	2.95 ☆☆☆☆	1285.63	2.95 ☆☆☆☆	1362.06	2.95 ☆☆☆☆	1363.28	2.95 ☆☆☆☆	1288.84	2.95 ☆☆☆☆	1288.94	2.90 ☆☆☆☆	1457.91	2.90 ☆☆☆☆	1457.94	2.95 ☆☆☆☆	1363.67	2.95 ☆☆☆☆	1363.67	2.96 ☆☆☆☆	1229.84	2.96 ☆☆☆☆	1229.84
Number of Authors	0.16 ☆☆☆☆	59.12	0.14 ☆☆☆☆	51.69	0.14 ☆☆☆☆	53.73	0.16 ☆☆☆☆	59.50	0.17 ☆☆☆☆	59.29	0.16 ☆☆☆☆	68.49	0.16 ☆☆☆☆	68.56	0.14 ☆☆☆☆	53,97	0.14 ☆☆☆☆	55.01	0.14 ☆☆☆☆	48.77	0.14 ☆☆☆☆	48.77
Number of References	0.29 ☆☆☆☆	116.70	0.24 ☆☆☆☆	101.05	0.24 ☆☆☆☆	100.35	0.28 ☆☆☆☆	111.60	0.28 ☆☆☆☆	111.55	0.30 ☆☆☆☆	140.96	0.30 ☆☆☆☆	140.97	0.24 ☆☆☆☆	98,38	0.24 ☆☆☆☆	98.33	0.24 ☆☆☆☆	89.90	0.24 ☆☆☆☆	89.91
Number of Institutes	0.05 ☆☆☆☆	15.61	0.03 ☆☆☆☆	11.29	0.03 ☆☆☆☆	11.75	0.05 ☆☆☆☆	15.37	0.05 ☆☆☆☆	15.30	0.04 ☆☆☆☆	16.83	0.04 ☆☆☆☆	16.85	0.03 ☆☆☆☆	11,72	0.03 ☆☆☆☆	11.50	0.03 ☆☆☆☆	10.54	0.03 ☆☆☆☆	10.53
Number of Countries	0.07 ☆☆☆☆	23.67	0.05 ☆☆☆☆	18.35	0.05 ☆☆☆☆	18.30	0.07 ☆☆☆☆	23.53	0.06 ☆☆☆☆	23.46	0.06 ☆☆☆☆	25.27	0.06 ☆☆☆☆	25.29	0.05 ☆☆☆☆	18,30	0.05 ☆☆☆☆	18.15	0.05 ☆☆☆☆	16.54	0.05 ☆☆☆☆	16.54
F-stastistic		243.6		380.2		380.4		248.0		246.7		332.7		330.9		379.0		377.9		0.30		0.30
R^2-adjusted		0.21		0.30		0.30		0.22		0.22		0.22		0.22		0.30		0.30		307.9		306.3

Table A.1: Robustness, no control for meshcount paper

☆ =$\rho < 0.1$; ☆☆ =$\rho < 0.05$; ☆☆☆ →$\rho < 0.01$; ☆☆☆☆ →$\rho < 0.001$;

All results are fixed for field variety

Field Names For Which The Results Are Fixed
ACOUSTICS
AUTOMATION & CONTROL SYSTEMS
AGRICULTURE, DAIRY & ANIMAL SCIENCE
AGRICULTURAL ECONOMICS & POLICY
AGRICULTURE, MULTIDISCIPLINARY
ENGINEERING, AEROSPACE
AGRICULTURAL EXPERIMENT STATION REPORTS
AGRONOMY
ALLERGY
ANATOMY & MORPHOLOGY
ANDROLOGY
ANESTHESIOLOGY
BIODIVERSITY CONSERVATION
ANTHROPOLOGY
ARCHAEOLOGY
ARCHITECTURE
AREA STUDIES
ART
HUMANITIES, MULTIDISCIPLINARY
ASTRONOMY & ASTROPHYSICS
PSYCHOLOGY, BIOLOGICAL
BEHAVIORAL SCIENCES
BIOCHEMICAL RESEARCH METHODS
BIOCHEMISTRY & MOLECULAR BIOLOGY
BIOLOGY
BIOPHYSICS
BIOTECHNOLOGY & APPLIED MICROBIOLOGY
PLANT SCIENCES
BUSINESS
BUSINESS, FINANCE
ONCOLOGY
CARDIAC & CARDIOVASCULAR SYSTEMS
CELL BIOLOGY
THERMODYNAMICS
CHEMISTRY, APPLIED
CHEMISTRY, MEDICINAL
CHEMISTRY, MULTIDISCIPLINARY
CHEMISTRY, ANALYTICAL
CHEMISTRY, INORGANIC & NUCLEAR
CHEMISTRY, ORGANIC
CHEMISTRY, PHYSICAL
CLASSICS
COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE
PSYCHOLOGY, CLINICAL

COMPUTER SCIENCE, CYBERNETICS
COMPUTER SCIENCE, HARDWARE & ARCHITECTURE
COMPUTER SCIENCE, INFORMATION SYSTEMS
COMMUNICATION
COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS
COMPUTER SCIENCE, SOFTWARE ENGINEERING
COMPUTER SCIENCE, THEORY & METHODS
CONSTRUCTION & BUILDING TECHNOLOGY
CRIMINOLOGY & PENOLOGY
EMERGENCY MEDICINE
CRYSTALLOGRAPHY
DANCE
DEMOGRAPHY
DENTISTRY/ORAL SURGERY & MEDICINE
DERMATOLOGY
GEOCHEMISTRY & GEOPHYSICS
SUBSTANCE ABUSE
ECOLOGY
ECONOMICS
EDUCATION & EDUCATIONAL RESEARCH
EDUCATION, SCIENTIFIC DISCIPLINES
EDUCATION, SPECIAL
PSYCHOLOGY, EDUCATIONAL
ELECTROCHEMISTRY
EVOLUTIONARY BIOLOGY
DEVELOPMENTAL BIOLOGY
ENDOCRINOLOGY & METABOLISM
ENERGY & FUELS
ENGINEERING, MULTIDISCIPLINARY
ENGINEERING, BIOMEDICAL
ENGINEERING, ENVIRONMENTAL
ENGINEERING, CHEMICAL
ENGINEERING, INDUSTRIAL
ENGINEERING, MANUFACTURING
ENGINEERING, MARINE
ENGINEERING, CIVIL
ENGINEERING, OCEAN
ENGINEERING, PETROLEUM
ENGINEERING, ELECTRICAL & ELECTRONIC
ENGINEERING, MECHANICAL
ENTOMOLOGY
ENVIRONMENTAL SCIENCES
ENVIRONMENTAL STUDIES
ERGONOMICS

ETHNIC STUDIES
FAMILY STUDIES
FILM, RADIO, TELEVISION
FISHERIES
FOLKLORE
FOOD SCIENCE & TECHNOLOGY
FORESTRY
GASTROENTEROLOGY & HEPATOLOGY
GENETICS & HEREDITY
GEOGRAPHY
GEOGRAPHY, PHYSICAL
GEOLOGY
GEOSCIENCES, MULTIDISCIPLINARY
GERIATRICS & GERONTOLOGY
HEALTH POLICY & SERVICES
HEMATOLOGY
HISTORY
HISTORY & PHILOSOPHY OF SCIENCE
HISTORY OF SOCIAL SCIENCES
HORTICULTURE
PSYCHOLOGY, DEVELOPMENTAL
PUBLIC, ENVIRONMENTAL & OCCUPATIONAL HEALTH
IMMUNOLOGY
INDUSTRIAL RELATIONS & LABOR
INFECTIOUS DISEASES
PSYCHOLOGY, APPLIED
INFORMATION SCIENCE & LIBRARY SCIENCE
INSTRUMENTS & INSTRUMENTATION
INTERNATIONAL RELATIONS
LAW
MEDICINE, LEGAL
ASIAN STUDIES
LINGUISTICS
LIMNOLOGY
LANGUAGE & LINGUISTICS THEORY
LITERARY REVIEWS
LITERATURE
MANAGEMENT
LITERATURE, AFRICAN, AUSTRALIAN, CANADIAN
OPERATIONS RESEARCH & MANAGEMENT SCIENCE
LITERATURE, AMERICAN
LITERATURE, BRITISH ISLES
LITERATURE, GERMAN, DUTCH, SCANDINAVIAN
MARINE & FRESHWATER BIOLOGY

MATERIALS SCIENCE, PAPER & WOOD
MATERIALS SCIENCE, CERAMICS
MATERIALS SCIENCE, MULTIDISCIPLINARY
MATHEMATICS, APPLIED
MATHEMATICS, INTERDISCIPLINARY APPLICATIONS
MATHEMATICS
SOCIAL SCIENCES, MATHEMATICAL METHODS
MEDICAL INFORMATICS
MECHANICS
MEDICAL LABORATORY TECHNOLOGY
MEDICINE, GENERAL & INTERNAL
METALLURGY & METALLURGICAL ENGINEERING
MEDICINE, RESEARCH & EXPERIMENTAL
LITERATURE, ROMANCE
LITERATURE, SLAVIC
MATERIALS SCIENCE, BIOMATERIALS
MATERIALS SCIENCE, CHARACTERIZATION & TESTING
MATERIALS SCIENCE, COATINGS & FILMS
MATERIALS SCIENCE, COMPOSITES
MATERIALS SCIENCE, TEXTILES
MEDIEVAL & RENAISSANCE STUDIES
METEOROLOGY & ATMOSPHERIC SCIENCES
MICROBIOLOGY
MICROSCOPY
MINERALOGY
MULTIDISCIPLINARY SCIENCES
MUSIC
MYCOLOGY
CLINICAL NEUROLOGY
NEUROSCIENCES
NUCLEAR SCIENCE & TECHNOLOGY
NURSING
NUTRITION & DIETETICS
OBSTETRICS & GYNECOLOGY
OCEANOGRAPHY
REMOTE SENSING
OPHTHALMOLOGY
OPTICS
ORNITHOLOGY
ORTHOPEDICS
OTORHINOLARYNGOLOGY
PALEONTOLOGY
PARASITOLOGY
PATHOLOGY

PEDIATRICS
PHARMACOLOGY & PHARMACY
PHILOSOPHY
PHYSICS, APPLIED
IMAGING SCIENCE & PHOTOGRAPHIC TECHNOLOGY
PHYSICS, FLUIDS & PLASMAS
PHYSICS, ATOMIC, MOLECULAR & CHEMICAL
PHYSICS, MULTIDISCIPLINARY
PHYSICS, CONDENSED MATTER
PHYSIOLOGY
PHYSICS, NUCLEAR
PHYSICS, PARTICLES & FIELDS
PLANNING & DEVELOPMENT
PHYSICS, MATHEMATICAL
POETRY
POLITICAL SCIENCE
POLYMER SCIENCE
PSYCHIATRY
PSYCHOLOGY, MULTIDISCIPLINARY
PUBLIC ADMINISTRATION
PSYCHOLOGY, PSYCHOANALYSIS
PSYCHOLOGY, MATHEMATICAL
PSYCHOLOGY, EXPERIMENTAL
RADIOLOGY, NUCLEAR MEDICINE & MEDICAL IMAGING
REHABILITATION
RESPIRATORY SYSTEM
REPRODUCTIVE BIOLOGY
RHEUMATOLOGY
SOCIAL ISSUES
PSYCHOLOGY, SOCIAL
SOCIAL SCIENCES, INTERDISCIPLINARY
SOCIAL SCIENCES, BIOMEDICAL
SOCIAL WORK
SOCIOLOGY
SOIL SCIENCE
SPECTROSCOPY
SPORT SCIENCES
STATISTICS & PROBABILITY
SURGERY
TELECOMMUNICATIONS
THEATER
RELIGION
TOXICOLOGY
TRANSPLANTATION

TRANSPORTATION
TROPICAL MEDICINE
URBAN STUDIES
UROLOGY & NEPHROLOGY
VETERINARY SCIENCES
PERIPHERAL VASCULAR DISEASE
VIROLOGY
WOMENS STUDIES
ZOOLOGY
MINING & MINERAL PROCESSING
WATER RESOURCES
ETHICS
HOSPITALITY, LEISURE, SPORT & TOURISM
HEALTH CARE SCIENCES & SERVICES
TRANSPORTATION SCIENCE & TECHNOLOGY
LITERARY THEORY & CRITICISM
AGRICULTURAL ENGINEERING
CRITICAL CARE MEDICINE
MATHEMATICAL & COMPUTATIONAL BIOLOGY
ENGINEERING, GEOLOGICAL
INTEGRATIVE & COMPLEMENTARY MEDICINE
NEUROIMAGING
GERONTOLOGY
ROBOTICS
NANOSCIENCE & NANOTECHNOLOGY
CULTURAL STUDIES
MEDICAL ETHICS
CELL & TISSUE ENGINEERING
PRIMARY HEALTH CARE
AUDIOLOGY & SPEECH-LANGUAGE PATHOLOGY
LOGIC
GREEN & SUSTAINABLE SCIENCE & TECHNOLOGY

Table A.2: All fieldnames which were used as dummy variables