UC Merced
UC Merced Electronic Theses and Dissertations

Title
Topic modeling in scientometrics: Community, connectivity, and change

Permalink
https://escholarship.org/uc/item/0zm8h881

Author
Bergmann, Till Christian

Publication Date
2016

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, MERCED

Topic modeling in scientometrics: Community, connectivity, and change

A dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Cognitive and Information Sciences

by

Till Christian Bergmann

Committee in charge:

Professor Teenie Matlock, Co-Chair
Professor Rick Dale, Co-Chair
Professor Michael Spivey
Professor Harish S. Bhat

2016

Portions of Chapter 3 © Till Christian Bergmann, Rick Dale

All other chapters © 2016 Till Christian Bergmann

All rights reserved

The dissertation of Till Christian Bergmann is approved,
and it is acceptable

in quality and form for publication on microfilm and electronically:

Professor Teenie Matlock, Co-Chair

Professor Rick Dale, Co-Chair

Professor Michael Spivey

Professor Harish S. Bhat

University of California, Merced

2016

To my parents,
who always supported my endeavors,
wherever they led me;

To my friends,
all over the world,
who put up with my cynicism;

And to Kelley,
without whom this dissertation
would not exist.

Contents

List of Figures

List of Tables

Acknowledgements

Later in this dissertation, I will argue that scientific publications are the result of their environment and influenced by their social network. It is therefore not a surprise that my dissertation is similarly the product of my professional and social environment. The Cognitive and Information Sciences department at UC Merced has provided an environment in which graduate students are encouraged to pursue their own research independently, and I would like to thank all members of the department for their support and feedback throughout the years. I would also like to thank the department for providing exceptional financial support, allowing me to concentrate on my research.

In particular, I would like to thank my advisors, Teenie Matlock and Rick Dale. Teenie has always believed in my academic potential and allowed me to pursue my ever-changing interests. Rick was not only a great mentor, but has also become a good friend. Thanks for all the discussions, work-related or not. Both have also provided me with excellent financial support beyond the departmental funding. I would also like to thank committee members Michael Spivey and Harish S. Bhat. Spivey has always encouraged me to consider the bigger cognitive picture, something that is often neglected in current research. Harish has rekindled my love for math(s), and is to blame for all equations in this dissertation. He has been a great help in forming the quantitative foundation of my work.

Thank you to all my friends, colleagues and collaborators who made my grad school experience what it was. Thanks to Chelsea, Collin, Dave, Justin, Katherine, Michelle and Spencer – even though some (most, actually) of you abandoned me in Merced – you have all been invaluable in keeping me (relatively) sane. Thanks to all my friends from the old continent, especially Anne, Faris, Leyla and Manu for keeping up with my exploits, and offering me asylum whenever I visit.

Thank you to my parents, who have always supported my quest for education and learning, no matter where it led me. And it usually led me abroad. You have been great role models in how to live a life, and have taught me all the values I cherish today.

Thank you to Keiko, Mike and Kristine for always taking me in and being so welcoming. The time with you has allowed me to relax, recharge, and appreciate that there is an "outside world" beyond academia.

My biggest gratitude is to Kelley. Without you, I would have never reached the end. You have been my biggest fan when I succeeded, and my biggest believer even when I failed. Words cannot express how much I owe you.

Chapter 3 is a partial reprint of a paper titled "A Scientometric Analysis Of EvoLang: Intersections And Authorships", co-authored by Rick Dale. Chapter 4 is a partial reprint of a manuscript titled "How Cognitive is Cognitive Linguistics? A Quantitative Analysis.", co-authored by Rick Dale and Teenie Matlock. Chapter 5 is a partial reprint of a manuscript titled "Comparing patterns of change in science and the humanities", co-authored by Rick Dale and Harish S. Bhat. I thank them for their continuous feedback during the analysis and writing stage, and all errors remain my own.

Curriculum Vitæ

Education

2016	PhD in Cognitive and Information Sciences
	University of California, Merced
2012	B.A. in English Language; Modern History
	University of Heidelberg, Germany

Peer-Reviewed Papers

Bergmann, T., Dale, R. & Lupyan, G. (in press). Socio-demographic influences on language structure and change: Not all learners are the same (commentary on Christiansen & Chater, 2015). *Behavioral and Brain Sciences*.

Bergmann, T., Dale, R., Sattari, N., Heit, E. & Bhat, H. S. (2016). The Interdisciplinarity of Collaborations in *Cognitive Science*. *Cognitive Science*.

Bergmann, T. & Dale, R. (2016). A Scientometric Analysis Of Evolang: Intersections And Authorships. In Roberts, S., Cuskley, C., McCrohon, L., Barceló-Coblijn, L., Feher, O. & Verhoef, T. (Eds.), *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANGX11)*.

Matlock, T. & **Bergmann, T.** (2015). Fictive Motion. In Dabrowska, E. & Divjak, D. (Eds.), *Handbook of Cognitive Linguistics*. 546-562. Amsterdam: DeGruyter Mouton.

Bergmann, T., Dale, R. & Lupyan, G. (2014). Informational structure of an emerging communication system is shaped by its environment. In Cartmill, E., Roberts, S., Lyn, H. & Cornish, H. (Eds.), *The Evolution of Language. Proceedings of the 10th International Conference (EVOLANG10)*. 387-388.

Bergmann, T., Dale, R. & Lupyan, G. (2013). The Impact of Communicative Constraints on the Emergence of a Graphical Communication System. In Knauff, M., Pauen, M., Sebanz, N. & Wachsmuth, I. (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. 1887-1992.

Talks and Posters

Heit, E., **Bergmann, T.**, Bhat, H. S. & Dale, R. (accepted). *A bibliometric appoach to studying group reasoning*. Talk to be presented at the International Conference on Thinking.

Padilla, L., **Bergmann, T.** & Creem-Regehr, S. (2016). *Uncertainty in Weather Forecast Phrasing*. Poster presented at the International Meeting of the Psychonomics Society.

Bergmann, T. (2016). *Data exploration, model diagnostics and visualization with R*. Presented at the 5th Conference on Statistical Practice 2016, San Diego.

Bergmann, T. & Matlock, T. (2015). *Watching Fictive Motion in Action: Discourse Data from the TV News Archive*. Poster presented at the 37th Annual Conference of the Cognitive Science Conference, Pasadena.

Bergmann, T. & Dale, R. (2015). *Quantifying academic research: A case study*. Presented at the Social Computing Working Group, D-Lab, UC Berkeley.

Coe, C., **Bergmann, T.** & Matlock, T. (2015). *Violence Metaphors in Presidential Debates*. Poster presented at the 37th Annual Conference of the Cognitive Science Conference, Pasadena.

Matlock, T., Gann, T., **Bergmann, T.** & Coe, C. (2015). *Metaphor in communicating wildfire risk*. Presented at the 2015 Conference on Communication and Environment, Boulder, Colorado.

Bergmann, T. (2014). *Observing Fictive Motion in the Wild*. Invited talk, CogNetwork, University of California, Berkeley.

Bergmann, T. & Matlock, T. (2014). *Fictive Motion and Gestures: Real Discourse Data from the TV News Archive*. Presented at ISGS 6, San Diego.

Bergmann, T. & Matlock, T. (2014). *Looking at fictive motion in natural discourse*. Presented at CSDL 2014, Santa Barbara.

Bergmann, T., Banks, C. & Matlock, T. (2014). *Watching fictive motion in action*. Presented at RaAM 10, Cagliary, Italy.

Matlock, T., Westerling, A. L., Gann, T., **Bergmann, T.** & Banks, C. (2014). *How we Talk about Wildfires*. Presented at the 94th American Meteorological Society Annual Meeting, Session: Ways of Speaking: The Role of Language and Culture in the Production, Communication, and Interpretation of Weather Information.

Bergmann, T., Dale, R. & Lupyan, G. (2014). *Informational structure of an emerging communication system is shaped by its environment*. Presented at EVOLANG 10, Vienna, Austria.

Bergmann, T., Dale, R. & Lupyan, G. (2013). *The Impact of Communicative Constraints on the Emergence of a Graphical Communication System*. Poster presented at the 35th Annual Conference of the Cognitive Science Conference, Berlin.

Bergmann, T. & Matlock, T. (2013). *Fictive motion in action: Gestures and visual representations co-occurring with fictive motion sentences in TV news*. Presented at ESLP 2012, Potsdam, Germany.

Bergmann, T. (2013). *The Historical Development of Fictive Motion*. Presented at the 12th International Cognitive Linguistics Conference (ICLC), Edmonton, Alberta.

Bergmann, T. & Pleyer, M. (2012). *Interdisciplinary Approaches to Construal Operations*. Presented at the 4th UK Cognitive Linguistics Conference, King's College London.

Abstract

This dissertation examines the complex structure of scientific organization and publication behavior. Since the last century, the number of scientific publications has exponentially risen, and researchers are now more connected than ever. This has led to an increasing interest in quantifying the structure of academia at various levels, for exampling university rankings and journal impact factors. In the current work, abstracts of scientific publication will be analyzed with respect to three different features of academia. First, the internal structure of a scientific community will be examined. What are the research topics prevalent in a community? Which are neglected? Does collaboration between researchers facilitate or hinder topic popularity? We find that more central authors in the community publish on a distinct set on research areas than non-central authors. Second, the connectivity between different scientific communities will be analyzed. Using quantitative methods, the overlap in scientific content between related scientific fields will be measured. Despite claims from within the community, the quantitative analysis shows very little overlap between supposedly related areas. Third, temporal change of scientific fields will be investigated. Taking two unrelated fields, philosophy and biology, the change of topics is evaluated over time. It is shown that biology as a field undergoes more change than philosophy, and the patterns of topic change differ across the two fields. In biology, topics either increase or decrease in popularity, while in philosophy their popularity fluctuates up and down over time.

These different aspects of scientific organization will be examined using topic models, a tool from natural language processing, and extended by various methods for each chapter. The theoretical discussion will argue that the results obtained in the case studies are heavily influenced by group cognition, that is, pressures and influences inherent to social groups.

This dissertation, *Topic modeling in scientometrics: Community, connectivity, and change*, is submitted by Till Christian Bergmann in 2016 in partial fulfillment of the degree Doctor of Philosophy in Cognitive and Information Sciences at the University of California, Merced, under the guidance of dissertation committee co-chairs Teenie Matlock and Rick Dale.

CHAPTER 1

Introduction

Science and academia are complex systems with many agents at different levels. At the micro-level, individuals pursue research and hold academic positions. At the mesa-level, these individuals form collaboration teams, departments and universities and pursue goals together. At the macro-level, the universities are linked together by global communication and collaborations. This complexity makes quantification and evaluation of these structures difficult. How are different universities linked together? How are different fields of science connected? Which individuals push the boundaries of science further? How should funding agencies allocate funding across this vast system? The field of *scientometrics* tries to answer such questions using quantitative, large-scale methods (Leydesdorff, 2001; Leydesdorff & Milojević, 2015). Increase in computation power and the rising availability of accessible databases of scientific information has made such an approach possible.

Researchers in scientometrics have looked at varying levels of organizational structure. For example, at the macro-level, attempts have been made to rank universities world-wide (Shin, Toutkoushian, & Teichler, 2011). Such rankings are not only important for universities for their prestige, but also affect how their research is perceived and can influence whether students will attend the university. At the mesa-level, scientometrics studies the relationship between different scientific fields, as well as the internal structure of those fields. Using citation data, Goldstone and Leydesdorff (2006) analyzed how the interdisciplinary field of cognitive science draws inspiration from other fields, and which fields in turn are likely to cite cognitive science, thus measuring which fields are influenced by cognitive science. Other approaches have relied on departmental affiliation of the authors (Gentner, 2010) and previous publication history (Bergmann, Dale, Sattari, Heit, & Bhat, 2016) to model the diversity of

a field. Studies have also identified which publications are especially impactful based on their citations and which years have shaped the current state of fields (Marx, Bornmann, Barth, & Leydesdorff, 2014; Wray & Bornmann, 2015). The pattern of papers cited in a paper is also tied to how well that paper is received, for example, both papers with high and low impact typically have more diverse citation patterns than papers with a medium impact, meaning they cite papers from multiple disciplines (Shi, Leskovec, & McFarland, 2010). Predicting impact of papers has also been modeled on the interdisciplinarity of the co-authors (Bhat, Huang, Rodriguez, Dale, & Heit, 2015), as well as using several features like the content of the paper and whether the authors are established in the community (Yan, Tang, Liu, Shan, & Li, 2011; Dong, Johnson, & Chawla, 2015). Using the content of papers, various papers have analyzed the scientific themes within fields and how they change over time (Hall, Jurafsky, & Manning, 2008; De Battisti, Ferrara, & Salini, 2015). Other historical analyses rely on author-provided keywords (Bentley, 2008). At the micro-level, traditionally the number of publications have been used as a measure for an individual's productivity, impact, and success. Large scale databases such as Google Scholar have made it possible to directly link citation counts to these publications, and attempts such as the *h*-index have been made to subsume citation counts into one single measure (Hirsch, 2005). As there is a continuing trend to publish in teams of authors, the micro-level does not play a huge rule in current scientometrics research, instead focus is given to how these teams of collaborators work together (Börner et al., 2010).

As evident from this list of studies, scientometrics uses a variety of variables as their measure of interest. Most of the variables are directly tied to scientific output and thus seem good choices for quantitative analysis, such as the abstract or title of papers, the list of authors and citation data. However, some variables are more constrained: Keywords are often restricted by journals to a certain set, and author department affiliation do not always accurately model the research they work on. For example, Gentner (2010) used departmental affiliation to argue that cognitive science is dominated by psychologists, but often universities do not have a separate cognitive science department, making the psychology department the most suitable affiliation for cognitive scientists, no matter what their exact research areas are within cognitive science. In this dissertation, abstracts of scientific publications are thus used as the basis of analysis, as they are arguably the richest data source and directly tied to the output of scientists.

Using the abstract data, the dissertation will analyze three different aspects of scientific organization through case studies:

1. What is the internal structure of a scientific *community*? Which different research areas are represented, and how are they related to the status of the authors working in these areas?
2. What is the *connectivity* between several scientific communities? Is there an overlap of scientific content between seemingly related fields?
3. How do scientific fields *change* over time? Do certain fields change more than others, and do they change in different ways?

Although the questions posed above are quite different in their nature, their analysis all rely on one computational tool, *topic modeling*. Topic modeling is a suite of natural language processing algorithms that allow the automatic extraction of topics, or *gists*, of large numbers of documents. Instead of having to go through documents manually, topic modeling can be used to automatically analyze them and represent them as a mixture of topics, where topics are defined a set of semantically related words. Chapter 2 will give an in-depth introduction to topic modeling, and how to use it to analyze scientific abstracts. It will explain the conceptual motivation behind topic modeling, the mathematical implementation of the inference mechanism, and features code-examples on how to run topic modeling in the programming language R (R Core Team, 2016).

The subsequent three chapters will use topic modeling to answer the three questions outlined above. Chapter 3 investigates the first question, the internal structure of a scientific community: Which kinds of topics are discussed within the community? Which topics are important within this community? Which topics are more to the periphery? Using the EvoLang (*Evolution of Language International Conference*) conferences as a case study, topic modeling will be combined with a social network analysis of collaborators to answer these questions. This combination allows to not only infer the topics present at the conference, but also how they are related to individual authors and author teams. The results indicate that certain topics are overrepresented within the scientific community, while others are neglected. Importantly, these results are partly incongruent with the posited goal of interdisciplinarity of the EvoLang community. Such analyses can thus help community to evaluate their current state, and whether research on certain topics needs to be encouraged.

Chapter 4 uses topic modeling to look at scientific connectivity across different fields. How are sub-fields within cognitive science related? Specifically, what is the relation between cognitive linguistics to general linguistics and

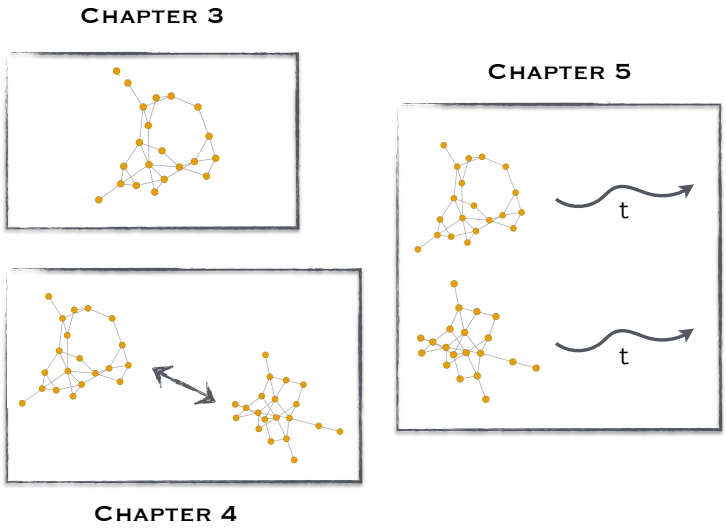

Figure 1.1: Chapter 3 looks at internal ties within a community, here depicted by the orange network. Chapter 4 looks at relationships between such scientific communities, while Chapter 5 studies the change in communities over time. Scientific communities are represented by orange networks.

cognitive science? Claims from cognitive linguists posit that cognitive linguistics is more related to cognitive science than other linguistics journals. Using topic modeling, the overlap between the output of different journals will be analyzed and measured, quantifying the subjective claims by members of the cognitive linguistics community. The topic model analysis is extended by a analysis of citation patterns, examining the relationship between the content of papers and their references. For example, does citing more cognitive science papers result in content that is more similar to the content of cognitive science papers?

Lastly, Chapter 5 uses topic modeling to quantify scientific change over time. By using historical data going back to 1980, the popularity of topics can be tracked through time with spline regression. The current case study looks at two distinct fields, biology and philosophy, which differ greatly in their subject matter and internal organization. The results indicate that topic models are a suitable way to track popularity of research themes over time, and furthermore, that topics in biology change more than in philosophy. The changes also show that topics in biology either rise or fall in popularity, representing topic overhaul, while topics in philosophy both rise and fall at different times, meaning that topics in philosophy are more akin to trends disappearing and reappearing.

Figure 1.1 shows the different aspects of scientific communities analyzed in the chapters. Scientific communities here are represented by the social networks in orange. Each chapter looks at a different dimension of these structures. Chapter 3 focuses on the internal structure, Chapter 4 at differences and similarities between communities, and Chapter 5 at internal changes over time. The use of topic models in all the analyses underlines its flexibility and extensibility, as it can easily be combined with other methods.

Lastly, the discussion chapter (Chapter 6) will tie the findings of the previous case studies together, and relate the findings of each case study to the "big picture" of collective behavior in academia. The case studies both reveal insight into the communities analyzed, but also relate to the more general study of how scientific communities interact and organize themselves. The qualitative analyses of the communities in Chapter 3 and Chapter 4 can help the communities to identify internal problems, and can lead to re-structuring and re-orienting of future research goals. Chapter 5 shows that different scientific fields change and progress differently, which affects how scientific work should be evaluated. Beyond the communities, the studies also provide support for the use of topics models in scientometric research, as it allows to study a wide range of academic aspects. The discussion chapter will also argue that a framework from cognitive science, *group cognition*, is one of the driving factors behind the results in the case studies. Scientists do not work in isolation and are part of a network at multiple scales. Graduate students work closely with their advisors, who are influenced by their colleagues and the bigger scientific community they are part of. It is argued that the increasingly interconnected network of scientists has a profound effect on the research areas scientists work on and which areas receive the most attention within the community.

The dissertation thus contributes to the fields of information science, cognitive science and scientometrics and in several ways:

1. It makes use of one single computational method, topic modeling, to study different aspects of scientific structure and shows that valuable information can be gained through it.
2. It provides insight for members of the scientific communities analyzed in the case studies, allowing communities to evaluate their current state of affairs.
3. It quantitatively shows that humanities differ from hard sciences in the way they change over time, suggesting that the fields should be evaluated in different ways, an aspect often neglected in current scientometrics work.

4. It argues for a unifying framework to study the behavior of scientists, filling a hole in current scientometric research which lacks theoretical explanation behind their findings.
5. It provides a "playground" of real world data to study group cognition, which traditionally has relied on agent based models or experiments.

Before continuing to the case studies, the next section will explain topic modeling in more detail and lay the methodological groundwork the dissertation.

CHAPTER 2

An introduction to topic modeling

2.1 Introduction

Advances in computational power and new techniques have made it possible to analyze large collections of texts automatically. Such large collections of text include newspaper articles going back a century, customer reviews on products or restaurants, and, in the case of this dissertation, scientific articles. The digitalization of such data has brought the opportunity to study these data in more detail, but the complexity and largeness of the data has also made it impossible for humans to sift through each document manually.

In the 1990s, first advances were made to detect latent semantic similarities between words based on their occurrences in documents. Latent Semantic Analysis (LSA, also Latent Semantic Indexing, LSI) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Landauer, Foltz, & Laham, 1998) converts raw documents into document-term-matrices, in which cells represents the frequency a certain term (row) in a given document (column). An example document-term-matrix can be constructed from the two documents below:

1. *The dog is chasing the cat.*
2. *The cat is running away from the dog. The cat runs around the corner.*

When dealing with actual data, the documents are of course much longer, and do not just consist of one or two sentences. Table 2.1 shows the document-term-matrix for these two documents, after words have been stemmed (e.g. "running" becomes "run") and punctuation removed. As both words and documents are now represented as vectors, document-term-matrices can be used to compare the similarity between words, as well as between documents by taking different similarity measures such as cosine similarity. However, as your dataset

grows, so does your document-term-matrix, and it becomes increasingly sparse: A lot of words will only occur in few documents, resulting in a lot of zero cells. Thus, LSA uses singular value decomposition (SVD) to reduce the dimensions, and find the dimensions at which the variability is greatest.

Table 2.1: Example of a document-term-matrix

	around	away	cat	chase	corner	dog	from	run	the
Doc1	0	0	1	1	0	1	0	0	2
Doc2	1	1	2	0	1	1	1	2	4

One of the drawbacks of LSA is the lack of transparency in similarity assessments: While it is possible to get a numeric representation of how similar two documents or terms are, it is not possible for humans to assess *why*. It is therefore not particularly useful when trying to summarize documents automatically, and it is more useful for comparisons between documents.

Topic modeling tries to alleviate this and represent documents as a mixture of *topics* that make up the *gist* of the document. The basic underlying idea is that each document comprises multiple topics, and each word in a document is assigned a topic. Each topic is thus a distribution over terms, that is, a distribution that expresses how likely it is for a word to occur in that topic. An example would be articles in a newspaper. Articles in newspapers cover a range of different topics, such as economics and sports. Topic models assume that each article is biased to talk about certain topics, rather than the whole range of topics. For example, an article in the sports section might talk about championships and winning, while an article in the economy section will rarely talk about these topics. In turn, such topics are also biased to include different terms. A topic about championships will be biased to include terms such as ring or cup, compared to words such as income or budget, which are more likely to occur in an economics topic.

Instead of a document-term-matrix (reduced via SVD), LDA represents documents as a probability distribution over topics.[1] For example, a document in the sports section might consist of 90% sports topic, but also 10% economy. Another document in the sports section might deal mainly with players' salary, and thus the distribution might be more evenly 50/50. Table 2.2 shows a hypothetical representation of documents as a mixture of T topics. The topics themselves are represent by a distribution over terms. Topic 1, sports, will give higher probability to sports terms than topic 2, about economy. An example distribution is shown in Table 2.3. These matrix representations allows us to

[1] In this work, we will use the word probability distribution to mean probability density function.

perform computations on documents, such as comparing similarities between documents and calculating popularity of topics over all documents, as well as succinctly summarize documents by their most used topics. Topic models thus have more flexibility and transparency in analyzing documents than LSA and other methods.

Table 2.2: Example topic distribution by document. Each cell represents the probability of a given topic (column) present in a given document (row).

	Topic 1	Topic 2	...	Topic T
Doc1	0.8	0.19	...	0.01
Doc2	0.1	0.89	...	0.01

Table 2.3: Example of term distribution topic. Each number expresses the probability that the term occurs under the topic. These probabilities differ across topics.

Topic 1 (sport)		Topic 2 (economy)	
championship	0.3	income	0.3
ring	0.15	budget	0.2
victory	0.1	salary	0.15
salary	0.08	trade	0.1

2.2 Latent Dirichlet Allocation

2.2.1 Generation of documents

While different types of topic models exist, we will concentrate on the most simple, and earliest type, Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). LDA automatically infers the topic and term distribution seen in the section above by probabilistic inference. Before the method of inference will be described in more detail, the generative aspect of LDA will be explained as it is conceptually helpful. The generative aspect means that new data can be generated once all parameters have been learned.

As mentioned above, in LDA each document is represented by a mixture of topics - some topics will be highly represented, some a little bit, and others not at all. More formally, a topic is a distribution over words. These topics are not

defined by topic modeling a-priori, instead they are automatically extracted based on the distribution of words in the corpus. The topics are post-hoc interpreted and labeled by the researchers, but as will be shown later, a topic model that fits the data well will allow such interpretation easily. A *sports* topic might include words such as championship, basket and assist to a high degree, while an *economics* topic is biased towards words such as income, budget and salary. LDA assumes that a document is generated in the following way:

1. Randomly sample a distribution over topics.
2. For each word in the document:
   (a) Randomly choose a topic from the distribution generated in the first step.
   (b) Randomly choose a word/term from the corresponding distributions over the vocabulary.

This generative process can be illustrated by a toy example. Figure 2.1 shows an example of term distributions for the two topics, *economy* and *sports*. The probabilities differ across the topics. For the generation of a new document, we start by assigning it a topic probability distribution. For example, it can consists of 80% sports topic, and 20% economy. For each word in the document, first, a topic is chosen from the prior topic distribution (in this case, 80% sports, 20% economy), and then a word from that topic given the prior term distribution (illustrated in Figure 2.1).

The technical details of topic modeling can be conveyed concisely in the following way. Both the term distribution ϕ and topic distribution θ are drawn from a Dirichlet distribution, $\theta \sim Dir(\alpha)$, $\phi \sim Dir(\beta)$. A Dirichlet distribution is a continuous multivariate probability distribution, which is commonly used in Bayesian statistics. In other words, a Dirichlet distribution represents a distribution over multinomial distributions. Step 1 of the above procedure translates to sampling one of these multinomial distributions from the Dirichlet prior α. For example, with the number of topics $T = 20$ and $\alpha = 0.1$, the resulting topic distribution θ_d for a document d takes the shape as depicted in Figure 2.2. Each topic (on the x-axis) is given a probability value (shown on the y-axis). The parameter α controls the shape of this distribution. A high value means that all values within the resulting multinomial θ_d are close to the mean, while a lower value increases the variance (Figure 2.3).

For the second step, each word w is assigned a topic z_w from the topic distribution θ_d. In the example, there is a high probability that the assigned topic will be topic 4 ($p_{z=4} = 0.058$), while there is only a low probability for topic

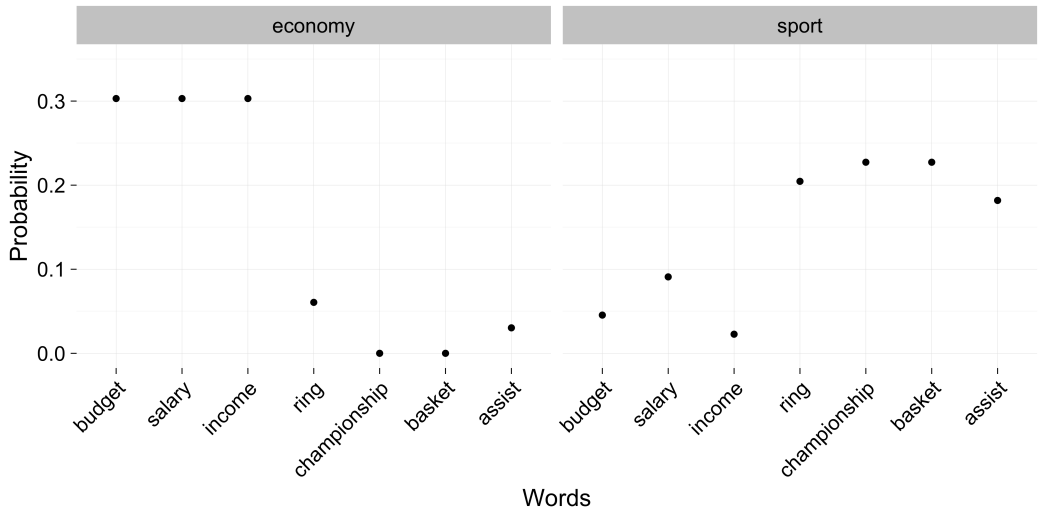

Figure 2.1: Sample term distributions by topic.

1 ($p_{z=1} < 0.0001$). After having drawn a topic z_w, a term is sampled from the similarly generated multinomial ϕ_z over all words in the vocabulary. An example of ϕ was shown previously in Figure 2.1.

The steps for generating a document with LDA are now the following:

1. For each topic z (where z is from 1 to T) generate a multinomial term distribution ϕ_z from a Dirichlet prior β to represent which terms are probable in which topics.
2. For each document d, draw a multinomial topic distribution θ_d from a Dirichlet prior α to represent which topics are probable in this document.
3. For each word w_{di} in document d:
   (a) Draw a topic z_{di} from θ_d
   (b) Draw a word w_{di} from $\phi_{z_{di}}$

Of course, when dealing with real data, documents do not need to be generated, instead, we want to infer the posterior distributions θ and ϕ. The next section will deal with the problem of turning the steps around, and inferring the posteriors.

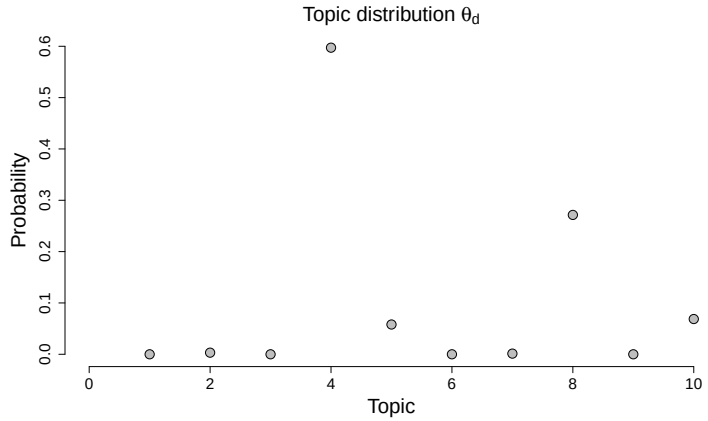

Figure 2.2: A topic distribution θ_d for a document d drawn from a Dirichlet prior. The topic distribution is biased towards 4 and 8, which means the document is mostly generated from words under those two topics.

2.2.2 Inferring the posterior

Gibbs Sampling

For practical applications, instead of generating new documents, we want to find out the topic distribution θ and term distribution ϕ. That is, for any given document d, what are the topics expressed in that document, and which terms are used for these topics?

One approach, first formulated by Griffiths and Steyvers (2004), is to use Gibbs sampling, a common algorithm within the Markov Chain Monte Carlo (MCMC) family of sampling algorithms (Casella & George, 1992; Gelfand, 2000). Gibbs sampling is useful when it is difficult (or impossible) to draw samples from a joint distribution of multiple variables, but easy to draw samples from conditional distributions. Let us assume our model has two parameters, θ_1 and θ_2.[2] Under Bayesian inference, the parameter space for the variables is the joint distribution of the parameters, $P(\theta_1, \theta_2 \mid D)$, where D is the data provided. This joint distribution can take any form. One complex example is shown in Figure 2.4. When the number of parameters increase, this joint distribution becomes more and more complex, and is often impossible to solve analytically. The approach taken in Gibbs sampling is instead to sample from the conditional probability distributions $P(\theta_i \mid \theta_{j \neq i}, D)$. In the case of two parameters, the conditional probabilities are $P(\theta_1 \mid \theta_2, D)$ and $P(\theta_2 \mid \theta_1, D)$. In Figure 2.4, we can sample

[2]Using θ to stand for parameters is standard Bayesian notation, and not related to the topic distribution in this example.

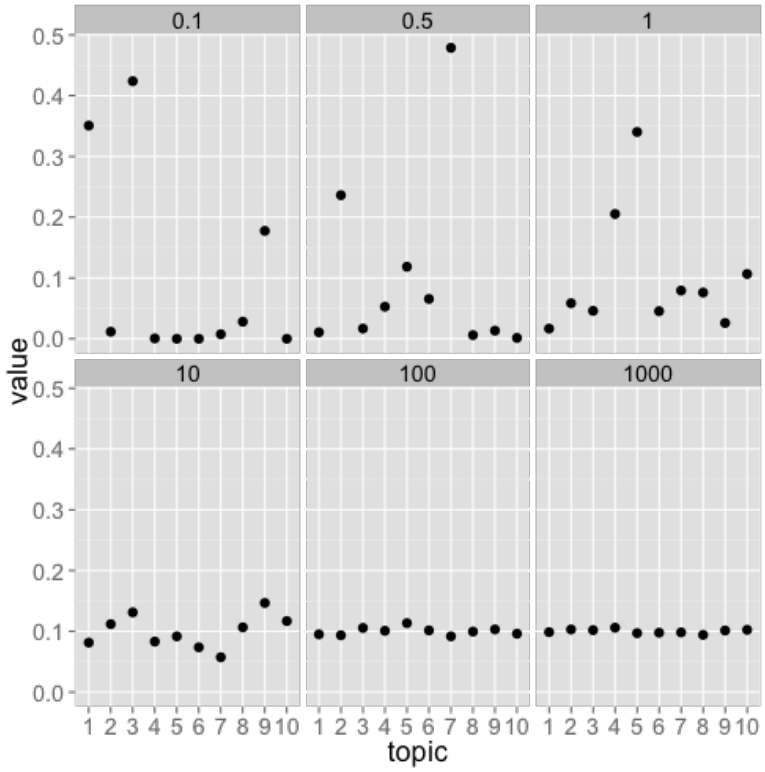

Figure 2.3: Different values of the Dirichlet parameter affect the distribution of the resulting multinomial. A low value (top left) results in a distributions where a few points are more probable than others, while a high value (bottom right) results in an even distribution. For this example, the number of topics T was set to 10.

θ_1 given a θ_2 value, and vice-versa. By iterating this process, the continious sampling from conditional probabilities equates to taking a random walk through the parameter space, where more time is spend in regions of the space that are more likely.

In each step of the Gibbs sampling procedure, a new value for a parameter is sampled according to its distribution conditioned on all *other* variables. This happens by cycling through all parameters sequentially. The updated values are immediately used as soon as they are updated. If there

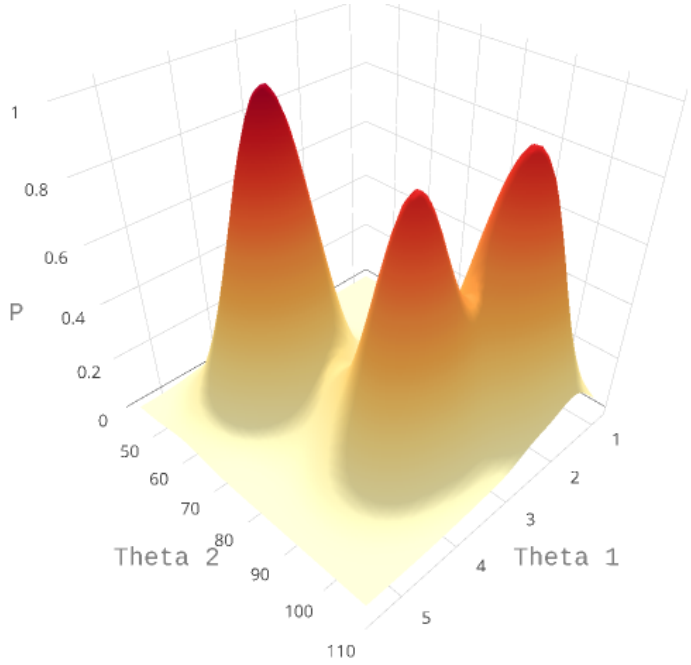

Figure 2.4: A complex joint probability distribution $P(\theta_1, \theta_2)$. The x- and y-axis express the values of θ_1 and θ_2, respectively, while the z-axis represents the probability density of those values. Gibbs sampling equates to taking a probabilistic random walk through this parameter space, spending more time in the regions that are more likely.

are three parameters, $\theta_1, \theta_2, \theta_3$, the algorithm can be summed up as follows:

1. Initialize $\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}$ to some value.
2. for each iteration i:
   (a) Draw a new value $\theta_1^{(i)}$ conditioned on values $\theta_2^{(i-1)}$ and $\theta_3^{(i-1)}$.
   (b) Draw a new value $\theta_2^{(i)}$ conditioned on values $\theta_1^{(i)}$ and $\theta_3^{(i-1)}$.
   (c) Draw a new value $\theta_3^{(i)}$ conditioned on values $\theta_1^{(i)}$ and $\theta_2^{(i)}$.

Gibbs sampling applied to LDA

Instead of finding estimates for the posterior distributions of θ and ϕ, Griffiths and Steyvers (2004) use an alternative approach of estimating the posterior distribution over the assignments of word tokens to topics, z, as both θ and ϕ can be calculated

from z. For each word token i, z_i is an integer value $[1 \dots T]$ representing the topic that it is assigned to. Once the topic assignment is known for each word token, we can easily calculate the distributions θ_d for each document and ϕ_j for each topic, as the words in each document are known.

Using Gibbs sampling, each document d_i and each word token in that document w_{di} is considered in turn, and its topic assignment z_i computed conditioned on the topic assignment on all other word tokens (Steyvers & Griffiths, 2007). In other words, the probability that a specific topic j is assigned to the current word w_{di} depends on the probability that the same word has been assigned that topic in other positions in the corpus. Formally, this posterior can be written as:

$$P\left(z_i = j | z_{-i}, w_i, d_i, \cdot\right) \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \quad \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha} \tag{2.1}$$

where $\cdot$ is all other known information, such as the Dirichlet priors and all other words w_{-i} and documents d_{-i}; and $\propto$ means proportional to, as in $y \propto x \equiv y = kx$.

C^{WT} and C^{DT} are matrices of counts with dimensions $W \times T$ (number of unique words in vocabulary $\times$ number of topics) and $D \times T$ (number of documents *times* number of topics) respectively:

- C^{WT}_{wj} is the count of word w assigned to topic j, not including current instance i.
- C^{DT}_{dj} is the count of of topic j assigned to some word token in document d not including current instance i.

Conceptually, the first ratio is the probability of w_i under topic j, and the second ratio the probability of topic j in document d_i. Once many tokens of word i have been assigned a topic j (across all documents), it will increase the probability that subsequent tokens of word i get the assignment topic j. Similarly, if topic j has been used multiple times *within* a document, it will increase the probability that any word within that document is assigned topic j.

Estimates of the topic distribution θ and term distribution ϕ can then be calculated using the following formula (Griffiths & Steyvers, 2004; Steyvers & Griffiths, 2007):

$$\theta_j'^{(d)} = \frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^{T} C_{dk}^{DT} + T\alpha} \tag{2.2}$$

$$\phi_i'^{(j)} = \frac{C_{ij}^{WT} + \beta}{\sum_{k=1}^{W} C_{kj}^{WT} + W\beta} \tag{2.3}$$

The Gibbs sampling procedure now can be written as:

1. assign each word token w_i a random topic $[1 \dots T]$
2. For each word token w_i:
   (a) Decrement count matrices C^{WT} and C^{DT} by one for current topic assignment.
   (b) Sample a new topic from equation 2.1.
   (c) Update count matrices C^{WT} and C^{DT} by one with the new sampled topic assignment.
3. Repeat above step *iter* times.
4. Calculate ϕ' and θ' from Gibbs samples z using equation 2.2 and 2.3.

Each Gibbs sample consists of a set of topic assignments to all N words in the corpus. There is an initial period, known as the *burn-in* period, where the samples are poor estimates and are usually discarded. After this period, the samples start to approach the target distribution. To get a representative sample, samples are saved at regularly spaced intervals to prevent correlation between them (a common problem in MCMC).

While Gibbs sampling has been commonly used in the inference step for LDA, other methods are possible. The original paper by Blei et al. (2003) uses variational inference, and Griffiths and Steyvers (2004) show that Gibbs sampling produces similar, if not more efficient, results. Alternative methods are also discussed in Blei and Lafferty (2009).

The use of Gibbs sampling to infer the posterior distributions also allows for easy extensions, for example, by including meta-information of documents. Meta-information such as the author of an article will influence the topic distribution, and this can be captured by including a hyper-parameter (Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004). Similarly, models exist to include the timestamp of documents (Wang & McCallum, 2006), as well as more general models to include any kind of meta-information (*structural topic model*; Roberts, Stewart, Tingley, and Airoldi, 2013).

Table 2.4: Documents generated with LDA from a rudimentary vocabulary.

d_i	Document Text
d_1	loan bank loan money money loan bank loan loan bank.
d_2	money loan money money money loan money money loan bank money.
⋮	⋮
d_{10}	river river bank stream bank bank bank bank stream bank.
d_{11}	stream bank stream stream stream bank stream bank.
⋮	⋮
d_{16}	money bank money loan loan bank loan bank money bank money bank.

2.2.3 Gibbs toy implementation

A toy implementation can be illustrated using an artifical example, where the topic and term distributions are known. Running the algorithm should result in similar distributions. Taking an example from Steyvers and Griffiths (2007), suppose that we only have two topics in our model, and our vocabulary only consists of five words, $V = money, loan, bank, river, stream$. Topic 1 gives equal probability to the first three words, i.e. $\phi^1_{money} = 1/3$, $\phi^1_{loan} = 1/3$, $\phi^1_{bank} = 1/3$, while topic 2 gives equal probability to the last three words, $\phi^2_{bank} = 1/3$, $\phi^2_{stream} = 1/3$, $\phi^2_{river} = 1/3$. Using these distributions, we can generate documents using the generative structure outlined above. For our example, we generated 16 documents, where each document has only been assigned one topic (rather than a mixture of topics) (Table 2.4). The first step is to randomly assign each word token w_i in the documents a topic assignment z_i. We can then calculate the count matrices C^{DT} and C^{WT}, shown below.

$$C^{DT}_{start} = \begin{bmatrix} 8 & 2 \\ 6 & 5 \\ \vdots & \vdots \\ 6 & 2 \\ 2 & 9 \\ \vdots & \vdots \\ 6 & 6 \end{bmatrix} \begin{matrix} d_1 \\ d_2 \\ \vdots \\ d_{10} \\ d_{11} \\ \vdots \\ d_{16} \end{matrix} \qquad C^{WT}_{start} = \begin{bmatrix} 11 & 14 \\ 20 & 7 \\ 29 & 28 \\ 11 & 17 \\ 18 & 17 \end{bmatrix} \begin{matrix} money \\ loan \\ bank \\ river \\ stream \end{matrix}$$

For the Gibbs sampling algorithm to work, we need a couple of pointers:

- i = index pointing to an individual word token.
- w_i = index pointing to the raw word in the vocabulary.
- d_i = index that tells you which document i belongs to.
- z_i = index that tells you what the topic assignment is for i.

After setting up these initial variables, we can start the Gibbs sampling process using the steps decribed in the above section. The Python code for the procedure is listed in Listing 1. Figure 2.5 shows how the probability of a word being assigned to a topic changes over 20 iterations. Even after a few iterations, the probabilities change according to the expected values. Figure 2.6 shows the original ϕ distribution on the left, the randomly initialized ϕ distribution at the beginning of the Gibbs sampling (center), and at the end (right). The Gibbs sampling procedure produces highly accurate estiamtes of ϕ in this example. Similar results are obtained for θ. For comparison to the original count matrices, the count matrices at the end of the sampling procedure show a much clearer picture:

$$C_{end}^{DT} = \begin{bmatrix} 10 & 0 \\ 11 & 0 \\ \vdots & \vdots \\ 0 & 8 \\ 1 & 10 \\ \vdots & \vdots \\ 12 & 0 \end{bmatrix} \begin{matrix} d_1 \\ d_2 \\ \vdots \\ d_{10} \\ d_{11} \\ \vdots \\ d_{16} \end{matrix} \qquad C_{end}^{WT} = \begin{bmatrix} 25 & 0 \\ 27 & 0 \\ 27 & 30 \\ 0 & 28 \\ 0 & 35 \end{bmatrix} \begin{matrix} money \\ loan \\ bank \\ river \\ stream \end{matrix}$$

2.3 Applying LDA in R

LDA using the Gibbs algorithm by Griffiths and Steyvers (2004) can easily be run using the *topicmodels* package in R (Grün & Hornik, 2011; Ponweiser, 2012) and the general text processing package *tm* (Feinerer, Hornik, & Meyer, 2008). The *tm* library also efficient text pre-processing tools such as stemming and removal of digits, stopwords and punctuation. Listing 2 shows a simple procedure in R for running LDA on documents. While the *topicmodels* package uses different labels for some of the variables, their purpose is the same. In the listing it is assumed that `documents` is a vector of pre-processed strings, where each element is a document.

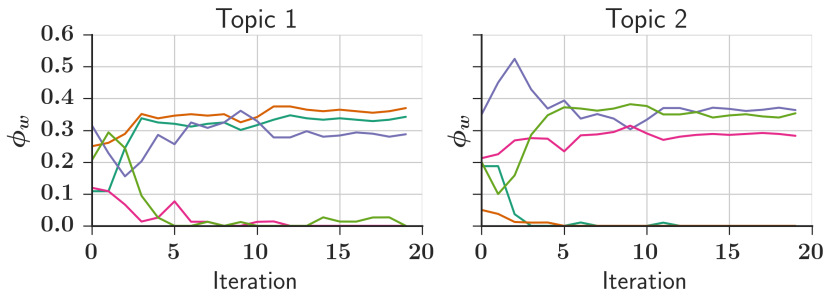

Figure 2.5: Changes in $\phi_w^{(j)}$ over iterations. Each line represents a word. Even after a few iterations, the probabilities change accordingly with the expected values.

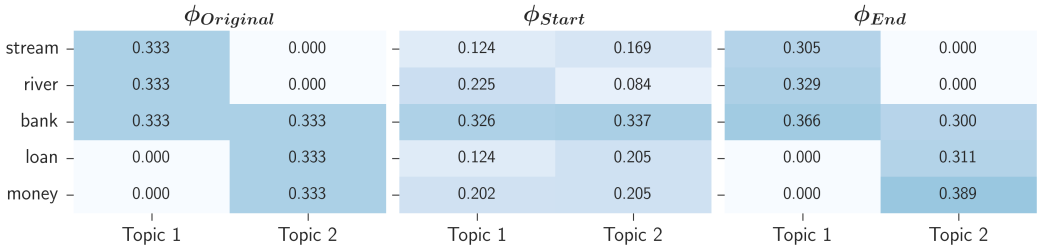

Figure 2.6: Visualizations of ϕ: The left matrix shows the original ϕ which was used to generate the documents. The center shows the randomly initialized ϕ at the start of the Gibbs sampling algorithm. The right matrix shows ϕ after 100 iterations of the Gibbs sampling procedure.

For further manipulation of parameters, please see the package documentation. After the model has successfully been fitted, the posterior distributions ϕ and θ can be examined using the commands `model@beta` and `model@gamma`, respectively (note again the diverging nomenclature).

```
import numpy as np

iters = 100
beta = 1.
alpha = 1.

for step in range(iters):
    # sample through each word
    for current in i:
        # get document index d_i
        doc_idx = d_i[current]
        # and word index w_i
        w_idx = w_i[current]

        # decrease count matrices C^DT and C^WT
        DT[doc_idx, z_i[current]] -= 1
        WT[w_idx, z_i[current]] -= 1

        # calculate new topic assignment z_i
        prob_word =  (WT[w_idx,:] + beta) /
                     (WT.sum(axis=0) + len(vocab)* beta)
        prob_document = (DT[doc_idx,:] + alpha) /
                        (DT.sum(axis=0) + D*alpha)
        prob = prob_word * prob_document

        # update z_i by sampling from the probabilities
        z_i[current] = np.random.choice([0,1], 1,
                                        p=prob/prob.sum())[0]

        # update count matrices C^DT and C^WT
        DT[doc_idx,z_i[current]] += 1
        WT[w_idx,z_i[current]] += 1
```

.

Listing 1: Gibbs sampling impsplemented in Python, using *numpy*. A complete working example can be found on Github, `https://github.com/tillbe/lda-gibbs-toy`.

```r
library(tm)
library(topicmodels)

# data
documents = c("words in document 1",
              "words in document 2",
              ...)

corpus = tm::Corpus(VectorSource(documents))
dtm = tm::DocumentTermMatrix(corpus)

# model parameters
T = 10 # number of topics
alpha = 50/T # dirichlet prior alpha
beta = 0.1 # dirichlet prior beta

model = topicsmodels::LDA(dtm,
                          k = k,
                          method = "Gibbs",
                          control = list(alpha=alpha,
                                         delta=beta)
                          )

# posteriors

# phi
model@beta

# theta
model@gamma
```

Listing 2: Running LDA in R with Gibbs Sampling using the *topicmodels* and *tm* packages.

CHAPTER 3

Structure within a scientific community

This chapter is an updated version of the following paper:
Bergmann, T. & Dale, R. (2016). A Scientometric Analysis Of Evolang: Intersections And Authorships. In Roberts, S., Cuskley, C., McCrohon, L., Barceló-Coblijn, L., Feher, O. & Verhoef, T. (Eds.), The Evolution of Language: Proceedings of the 11th International Conference (EVOLANGX11). Online at http://evolang.org/neworleans/papers/182.html.

Abstract: Research on the evolution of language has grown rapidly and is now a large and diverse field. Because of this growing complexity as a scientific domain, seeking new methods for exploring the field itself may help synthesize knowledge, compare theories, and identify conceptual intersections. In addition, it may help find gaps in the disciplinary composition of the area, in that some fields centrally related to human evolution may be surprisingly missing from conference presentations. Using computational methods, we analyze the scientific content presented at EvoLang conferences. Drawing on 565 abstracts, publication patterns are quantified using Latent Dirichlet Allocation (LDA), which extracts a semantic summary from individual abstracts. We then cluster these semantic summaries to reveal the frameworks and different domains present at EvoLang. Our results show that EvoLang is an interdisciplinary field, attracting research from various fields such as linguistics and animal studies. Furthermore, we show that the framework of iterated learning and cultural evolution is a hub topic at EvoLang.

3.1 Introduction

In this paper, we explore the conceptual structure of research on language evolution itself by analyzing the submissions to the EvoLang conference (*Evolution of Language International Conferences*, held bi-annually) over the past 10 years. Our goal is to provide insight into the network of theories, concepts, and methods that populate this growing field. Since its inception in 1996, EvoLang has become a prominent and well-attended conference. It is now the premiere conference on language evolution, with more than 100 presentations at the last EvoLang and over 300 delegates in attendance. This is a five-fold increase from the first EvoLang in 1996. How might we quantify this rapidly growing scientific content?

There are numerous reviews of language evolution which attempt to unpack and relate its various theories and debates (e.g. Christiansen & Kirby, 2003; Bickerton, 2007; Fitch, 2010). These provide impressive coverage, especially considering the diversity and complexity of language evolution research. Research at EvoLang tackles a wide range of these topics, spanning the many levels of language from the evolution of flexible signalling strategies to the social cognitive processes that may undergird human linguistic skills. Recently, the foundation of the *Journal of Language Evolution* aims to bring the wide range of research interests together, and provide an umbrella journal for language evolution research. Previously, researchers interested in language evolution published in a variety of journals, making it hard to keep track of all relevant publications. In an editorial article, the editors specifically address the interdisciplinarity of the research area (Dediu & de Boer, 2015). How much of this interdisciplinarity is already present at EvoLang, and what core areas relevant to language evolution do not yet have representation at the conference? Are there any approaches that are under- or over-represented? Using a quantitative approach, we can answer these questions.

In what follows, we use topic modeling (Chapter 2, Griffiths and Steyvers, 2004) to extract the set of latent conceptual topics that make up EvoLang. We find that there are three distinct conceptual clusters that can be inferred from the abstracts, including the iterated learning framework and comparative studies. Second, we combine these topic clusters with a co-authorship network analysis to assess the relative influence of these typical topic clusters, finding that the iterated learning cluster in particular serves as a central hub in the broader EvoLang community. By analyzing the knowledge bases of EvoLang, it may be possible to attain a firmer grip on the state of the art in the field, and the relationships among its various theories. Lastly, we will also briefly look at the evolution of

the conference over time, and show how these clusters change depending on the conference.

3.2 Modeling the content of EvoLang

3.2.1 Data and method

We selected all abstracts from submissions between 2006 and 2016 (conferences are bi-annual), and applied basic pre-processing to the abstract text. Abstracts before 2006 were published in a different format and were thus omitted to keep the data consistent. Pre-processing included tokenizing the text, removing punctuation and common words (stopwords) such as "the" and "or", and finally stemming the tokens using the Snowball stemmer (Porter, 2001). Abstracts with fewer than twenty stemmed tokens were removed from the analysis, as they did not provide enough information about the content of the paper, and manual inspection of these abstracts showed that they only contained the first sentences of the abstract and were cut off after.

We then applied Latent Dirichlet Allocation (see Chapter 2, Blei et al., 2003) on the resulting 565 abstracts, a method that is commonly used in scientific content analysis (Griffiths & Steyvers, 2004). In LDA, each document (here, abstract) is represented by a distribution over topics, and the topics themselves are represented by a distribution over words. That is, each topic consists of a distribution of semantically related words, and each abstract can then be represented as a combination of these topics, which make up the *gist* of the document. For example, one abstract at EvoLang may combine the topics of non-human communication and learning, while another may combine syntax and computation. Importantly, the algorithm only extracts numerically identified topics, and these hypothetical labels are assigned by the researchers. When the model fits the underlying data well, domain knowledge of the researchers combined with the associated words for each topic result in clear, intuitive labels. As we show below, this can result in a compelling set of topics.

3.2.2 Topics of EvoLang

After running the LDA algorithm with a various number of topics, we selected the model of best fit (based on log-likelihood), which contained 20 topics. Example topics are shown in Table 3.1 with associated terms. Note in the table that we have used a stemmer algorithm to obtain roots (e.g., "compar", "abil"), to decrease the type-token ratio, and facilitate topic extraction.

To further analyze the content and to investigate the relationships between these topics, a correlation matrix of the probability distributions for the topics was calculated and a network of positively related topics was generated. Then, a community detection algorithm (Pons & Latapy, 2005) was used to cluster these topics. We found that the algorithm clustered the content of EvoLang submissions broadly into three communities or clusters. The resulting network is shown in Figure 3.1, with the different clusters marked by color. Each node represents a topic, and each edge represents a positive correlation between two topics. Nodes are sized according to their overall popularity in the corpus, that is, larger nodes occur more often than smaller nodes.

But what do these clusters consist of? To get more insight into the topics associated with each cluster, we extracted the most probable terms associated with the topics in each cluster. The first cluster covers experimental research, including several topics on iterated learning and cultural evolution, as well as the emergence of structures in communication experiments (Table 3.1). The second cluster can be described as comparative studies involving primates and birds (Table 3.2). Lastly, papers in the third cluster approach language evolution through more traditional linguistic research, such as (universal) grammar (Table 3.3). Inspecting these terms and communities gives a good overview of different fields within EvoLang, and indeed, both the clustering and most probable terms make intuitive sense for researchers involved in the community.

In general, these clusters show that EvoLang hosts a variety of sub-fields, which approach the study of language evolution from varying angles. Not only does it include more theoretical linguistic work, but also comparative studies are well represented. Certainly this is well known intuitively by researchers within the community, but the analysis here suggests that there are crisp clusters that can be automatically extracted using the topic model. This suggests that the sub-fields of EvoLang either use a different set of words to talk about their research, or do not interact and collaborate to a great extend (or a combination thereof). In a recent editorial for the new journal *The Journal for Language Evolution*, Dediu and de Boer (2015) call for an interdisciplinary approach to language evolution based on sound empirical data. In the extracted topics, topic 2, 6, and 10 explicitly mention empirical methods, while topic 19 consists of computational modeling, another area mentioned by Dediu and de Boer (2015). However, other areas that make contributions to the study of language evolution are absent in the extracted topics. For example, genetics, which has made contributions in the field of genes involved in language such as *FOXP2*, is not represented at EvoLang, and neither

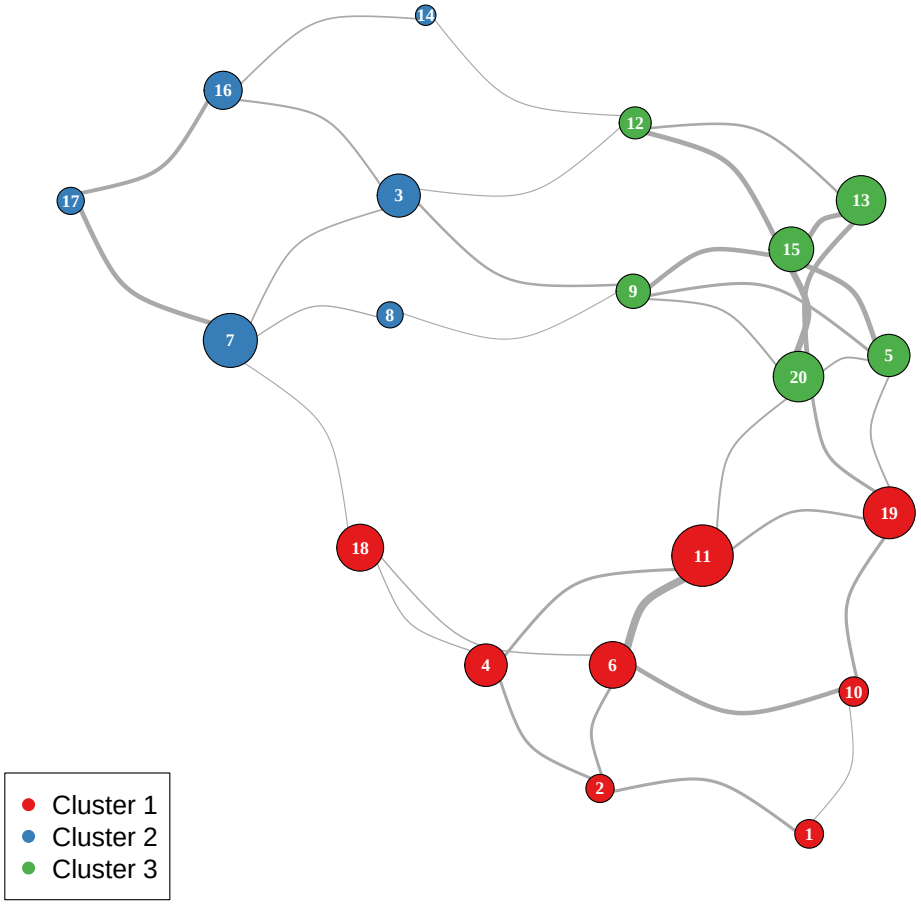

Figure 3.1: Network of positively correlated topics. The thicker an edge, the stronger the correlation. Node size represents topic popularity. The bigger a node, the more it is represented in the abstracts. Topics belonging to the same cluster share a color.

are anthropology and archeology. Despite the variety of sub-fields present in the topics, there are still gaps that can be addressed.

In the next section, we look at the author collaboration networks of EvoLang, and how they relate to these topic clusters. This serves both as an illustration of the range of authorship patterns, as well as being the measure through which we further analyze the interconnectedness of these three topic clusters.

Table 3.1: Topics in cluster 1 and their associated terms.

Topic 1	Topic 2	Topic 4	Topic 6
communic	word	mean	experi
inform	order	emerg	particip
speaker	product	languag	categori
relev	event	featur	studi
system	interpret	form	result
question	semant	composit	set
utter	data	space	condit
simpl	present	semant	task
encod	studi	combin	test
cue	lexicon	combinatori	present

Topic 10	Topic 11	Topic 18	Topic 19
signal	learn	gestur	model
game	cultur	languag	agent
system	bias	sign	popul
communic	structur	symbol	network
strategi	generat	system	interact
agent	languag	point	simul
interact	linguist	icon	communiti
high	regular	action	dynam
player	learner	speech	comput
refer	transmiss	form	effect

3.3 Modeling the authors of EvoLang

3.3.1 The topography of collaborations

By constructing an authorship network from co-authored abstracts, we can examine the nature of collaborations at EvoLang. Who collaborates with whom? What type of submission elicits large collaborations? Are there large components

Table 3.2: Topics in cluster 2 and their associated terms.

Topic 3	Topic 7	Topic 8	Topic 14	Topic 16	Topic 17
human	communic	brain	complex	vocal	social
develop	gestur	human	song	human	call
learn	human	involv	note	sound	individu
languag	ape	languag	finch	speech	group
mechan	primat	emot	speci	nonhuman	anim
abil	compar	studi	neural	primat	chimpanze
studi	intent	activ	increas	produc	time
cognit	signal	area	factor	acoust	behaviour
acquisit	research	origin	examin	vowel	level
stage	abil	relat	bird	speci	speci

Table 3.3: Topics in cluster 3 and their associated terms.

Topic 5	Topic 9	Topic 12	Topic 13	Topic 15	Topic 20
languag	evid	structur	cognit	evolut	languag
evolut	modern	syntact	process	languag	linguist
evolv	problem	grammar	system	process	argu
chang	protolanguag	rule	human	evolutionari	properti
behavior	hypothesi	syntax	role	natur	paper
explain	genet	recurs	evolut	biolog	univers
extend	homo	construct	capac	theori	origin
present	select	element	specif	faculti	question
work	make	acquir	framework	principl	term
spatial	potenti	determin	music	chomski	suggest

of connected collaborations? We can answer these questions by building a collaboration network from all EvoLang papers and their authors.

In this collaboration network, each node is an author and each edge between two nodes represents collaboration between these two nodes/authors. Edge weight (connection strength) is determined by the number of collaborations between these two authors. Using the topic clusters from the above analysis, we calculated the most prevalent cluster for each author. By aggregating the topic distributions across all papers by one author, the most common cluster was calculated for each author. By plotting the author network (Figure 3.2), we can see that there is one large hub in the middle of the network, as well as several smaller hubs of multiple nodes. Outside these hubs, a large quantity of small-scale collaborations exist, not connected to the rest of the network. These smaller collaborations often consist of advisor-advisee relationships within the same lab or department. The color of the nodes represents the respective cluster an author has mainly published in.

When we examine the local hubs more in detail, we notice that in the smaller components all nodes usually belong to one cluster. This intuitively makes sense, as such collaborations are usually just based on one or two papers, where all nodes are authors on the same papers and thus have the same distributions over topics and clusters. The larger hub, however, show more diversity – in the largest hub, all three clusters are present, although there is a dominance by cluster 1 (red). Figure 3.3 shows a close-up of the four components that have more than 30 nodes within them. Interestingly, while the largest component (top left), is dominated by cluster 1 (red), all other large components are almost exclusively about comparative and animal studies (blue). This shows that on the level of collaborations, this sub-field has its own sub-network of authors who frequently collaborate together, but is not connected to the central large component. At least on the author level, one can thus argue that there is a disconnect between these two groups, and authors from these two clusters (experimental work on cultural evolution, and comparative/animal studies) hardly collaborate together.

3.3.2 Centrality of authors and clusters

A network structure also allows a quantitative assessment which authors play a central role in the EvoLang community. Authors who publish and collaborate often are referred to as "central", and by virtue of their centrality, we can also assess the contribution of their associated topics in their collaborations. After constructing the network, centrality measures were used to detect the most influential authors within this network. In network theory, there are multiple

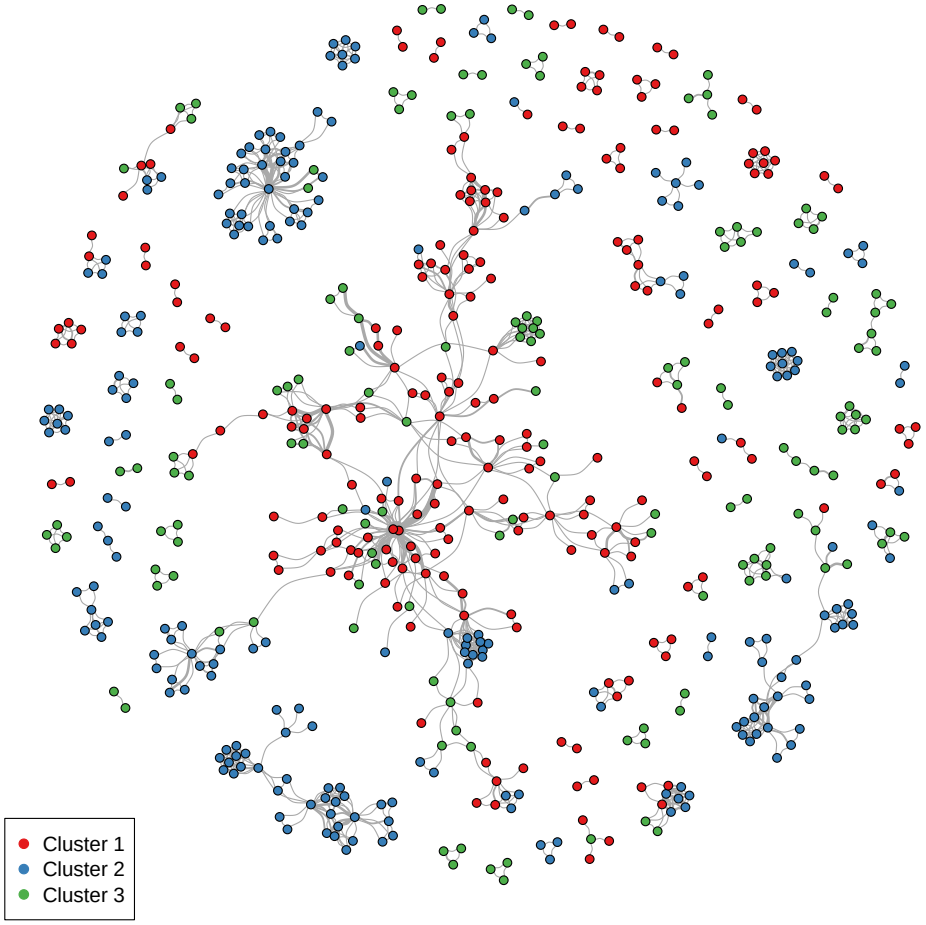

Figure 3.2: A network showing collaborations between authors. Nodes represent authors and are colored with respect to their dominant cluster. The thicker an edge, the more collaborations between the nodes.

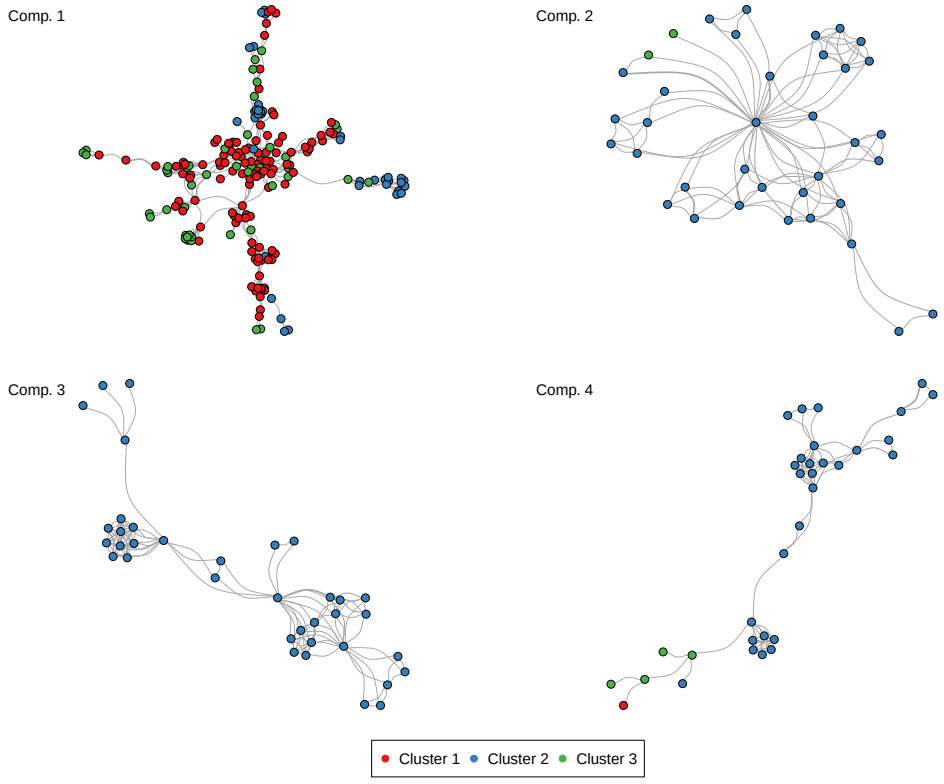

Figure 3.3: Four components/hubs with more than 30 nodes/authors. The largest component is dominated by Cluster 1 (green), with the other two clusters interspersed throughout, while the other three components are almost exclusively assigned Cluster 2, showing a strong sense of collaborations in animal studies.

ways to measure the centrality of nodes (Freeman, 1978; Koschützki et al., 2005; Kolaczyk, 2009). Here, we look at two values: eigenvector centrality and betweenness centrality. Eigenvector centrality measures the influence of a node by assigning a score based on connections to high scoring nodes (here, nodes with a lot of collaborations and thus submitted papers). The score is bound between 0 and 1, with 1 representing highest centrality. The equation for calculating eigenvector centraliy is:

$$c_{Ei}(v) = \alpha \sum_{u,v} c_{Ei}(u) \tag{3.1}$$

where the vector $\boldsymbol{c}_{Ei}$ is the solution to $\boldsymbol{A}\boldsymbol{c}_{Ei} = \alpha^{-1}\boldsymbol{c}_E i$, with $\boldsymbol{A}$ being the adjacency matrix for the graph.

Betweenness centrality assigns a score based on how often the node is part of the shortest path between two other nodes, and thus measures how well a node connects different parts of a network. Betweenness centrality is defined as:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{3.2}$$

where σ_{st} are all paths from node s to node t, and $\sigma_{st}(v)$ all are paths from node s to node t that go through node v. Nodes with a high value are considered to be important in communication between other nodes and keeping the network connected.

Fig. 3.4 shows the centrality measures of authors on a log-scale (purely for illustrating purposes): Authors with high eigenvector values but low betweenness have close contact to important people, while authors with low eigenvector values but high betweenness values serve as valuable connections between nodes. In the plot, there is a division between authors with a high and low eigenvector centrality. Authors with a high eigenvector centrality tend to be in cluster 1, while authors in cluster 3 are more likely to have low eigenvector centrality. Cluster 2 authors seem to be more interspersed. Authors to the right of the plot, with high eigenvector centrality, are part of the largest component ("Comp. 1"), and authors to the left are not. This results in their low eigenvector score, as they are not connected to the most central nodes as defined by eigenvector centrality.

By using the centrality measures calculated for each author, we were able to deduce the influence of each topic cluster. That is, to which cluster do the most widely collaborating individuals belong? Table 3.4 shows summary statistics for the author centrality measures in each cluster. Not surprisingly, cluster 3 has

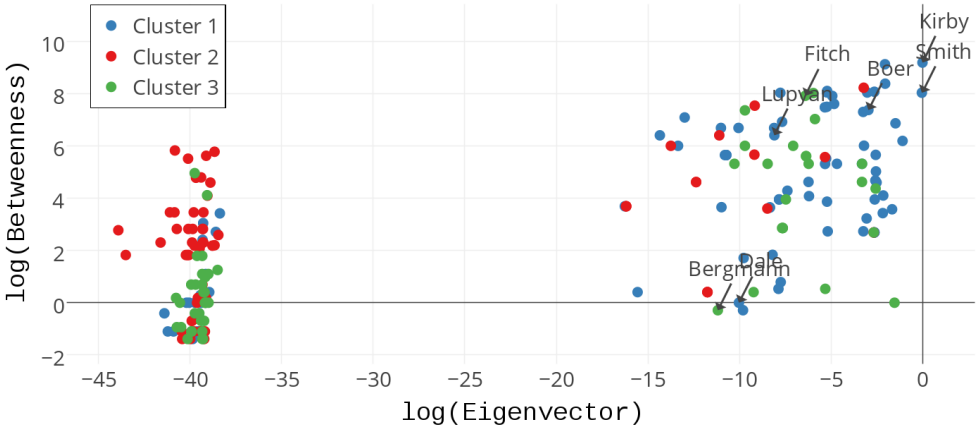

Figure 3.4: Betweenness and eigenvector centrality, on a log-scale. Each point represents an author, with the color representing their cluster. A few noteworthy authors are labeled.

Table 3.4: Summary statistics for each cluster of topics.

Cluster	M(Eigenvector)	SD(Eigenvector)	M(Betweenness)	SD(Betweenness)
1	0.0231	0.1017	304.12	0.1017
2	0.0009	0.0058	46.79	0.0058
3	0.0045	0.0232	76.09	0.0232

both the highest average eigenvector and betweenness centrality, however, it also has the highest deviations. While the deviations suggest that there is a lot of variation within clusters, it looks like cluster 3 is the most central set of topics within EvoLang.

To test whether this difference in centrality measures is significant, a multinomial logistic regression was run with the clusters as a dependent variable, and the two centrality measures as the independent measures. Cluster 1 was chosen as the baseline community, as we hypothesized that it had higher centrality than the other two clusters. The model output is summarized in Table 3.5 and was significant compared to a null model ($\chi^2(4) = 50.23, p < 0.0001$). Significance values were calculated using Wald tests. Coefficients for betweenness centrality were significant for cluster 2 ($p = 0.039$), but not for cluster 3 ($p = 0.1$). However, eigenvector centrality was a significant predictor for both clusters ($p < 0.0001$ for

Table 3.5: Summary of multinomial logistic regression showing log-odds and standard errors.

	Dependent variable:	
	Cluster 2	Cluster 3
Betweenness	-0.001^{**}	-0.0004
	(0.0003)	(0.0003)
Eigenvector	-43.651^{***}	-11.063^{***}
	(0.0001)	(0.001)
Constant	0.300^{***}	-0.225^{**}
	(0.101)	(0.112)

both clusters). As the log odds are very high, any increase in eigenvector centrality increases the probability of a paper being associated with cluster 1.

From this analysis, we conclude that cluster 1, which is strongly related to iterated learning and cultural evolution, serves as a "hub cluster" within EvoLang. The betweenness centrality showed a significant decrease for cluster 2 when compared to cluster 1. This means that authors in cluster 2 (comparative and animal studies) are more separated from the rest of the network, or in other words, they form their own sub-network to some degree. Authors in cluster 3, on the other hand, serve as an import connector between nodes in the network. Our hypothesis of a disconnect between authors in clusters 1 and 2 was thus confirmed.

3.4 Evolution over time

As data was present for all conferences from 2006 to 2016, it was also possible to analyze changes over time: Did clusters become more popular over time or decrease in their popularity? Did the same collaborations persist throughout the conferences, or do new combinations appear?

Figure 3.5 shows the percentage of papers in each respective cluster over each of the conference. In 2006, most of the papers belong to the third cluster, the linguistics cluster. Its popularity, however, slowly dwindles over time. It is replaced by cluster 1, the experimental cluster, which is steadily at the top, except in 2012, when the conference was held in Japan and attracted a different audience. The second cluster, about animal and comparative studies, has steadily increased in popularity. In 2016, all three clusters are very close to each other,

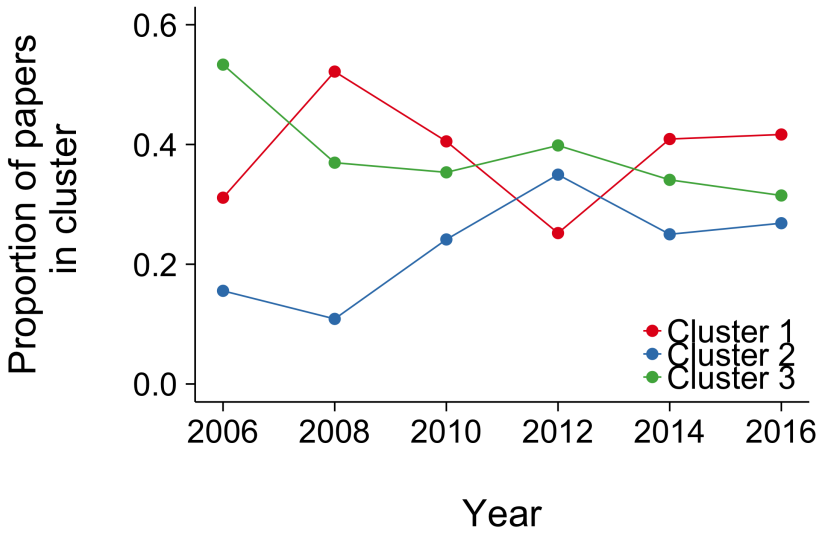

Figure 3.5: Proportions of papers in each cluster over time.

while before 2012, they were further apart. This is a sign that the conference has settled equally into these three clusters and that the extracted topic clusters are a good representation of the current state of EvoLang.

With regard to the author network, we plotted the collaboration network for each conference separately (Figure 3.6). The network in 2006 can hardly be described as such, it mainly consists of components of only two or three nodes. A similar picture is present for 2008. Beginning in 2010, the conference not only grows in size, but also in the scale of collaborations, and bigger components emerge. Starting in 2010, we have a large component in the center of the network, almost exclusively belonging to cluster 1. This hub persists in subsequent conferences, and grows even larger. In 2012, we also see a strong hub of cluster 2 nodes – as the conference was in Japan and the local university has a focus on animal studies, this makes intuitive sense.

The analysis over time has shown that the conference has grown in its collaborative nature, and seems to have settled into three clusters of research areas that are all equally represented. In the later years, the author networks looks very similar to each other, suggesting that the hubs and components persist over time in their nature.

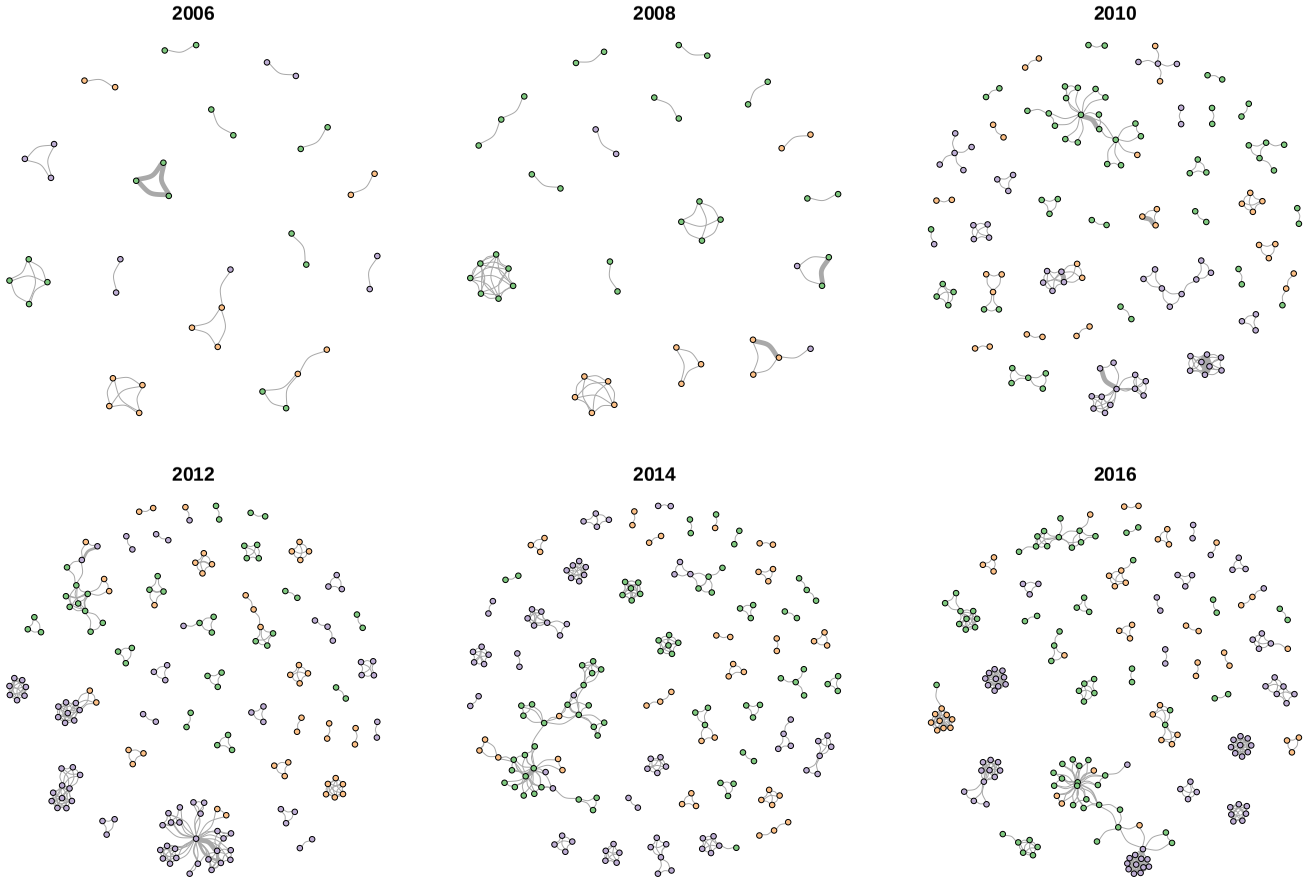

Figure 3.6: Evolution of network over time

3.5 Summary

We analyzed the content of abstracts presented at EvoLang. Our analysis of latent topics shows that EvoLang is an interdisciplinary conference, and draws attention from three major research topics. Despite the interdisciplinarity, we also identified areas that were under-represented at the conference, including work on genetics, archeology and anthropology. Using a network analysis of author collaborations, we investigated these clusters with regard to their influence. Our results suggest that the cluster containing the iterated learning and cultural evolution framework is associated with a high centrality property within EvoLang. Comparative studies with primates are rather disconnected from the rest of the network, and form their own sub-networks. Lastly, the cluster covering linguistic approaches is interspersed and well represented throughout the conference. The topic network shows that there seems to be a separation between certain research areas, and collaborations hardly include both animal studies and experimental design. However, this does not necessarily mean that ideas are not exchanged between these two groups, but raises concern how well the different research areas interact with each other. As will be elaborated in more detail in the General Discussion (Chapter 6), the structure of the social network of collaborators is very likely to directly influence the topics researchers work on. Filling in missing collaborations could thus lead to a more complete, holistic approach to language evolution.

Though these patterns may be intuitive to highly initiated attendees of the conference, the purpose of this paper is to demonstrate that scientometric techniques can be used to reveal these patterns quantitatively. With just under 600 abstracts, a number of natural authorship and conceptual patterns emerge. It may be useful and interesting to carry out similar analyses in subsequent years to discern how this field is changing, and how topic clusters may be converging or co-fertilizing.

CHAPTER 4

Structure across scientific communities

This chapter is an updated version of the following manuscript:
Bergmann, T., Dale, R. & Matlock, T. (manuscript). *How Cognitive is Cognitive Linguistics? A Quantitative Analysis*. Available online at `http://tillbergmann.com/diss/cogling_submitted.pdf`.

Abstract: Cognitive linguistics emphasizes its inherently interdisciplinary nature and its close link to the cognitive sciences. In this paper, we carry out quantitative analyses motivated by scientometrics to better understand the nature and extent of this interdisciplinarity, In the first section, the content of papers in *Cognitive Linguistics* (*CL*) is compared to the content of other journals using topic modeling. Our analysis shows that *CL* shares more content similarity with other linguistics journals than it does with cognitive journals, and more importantly, is not closer to cognitive journals than other linguistics journals are. A second analysis looks at the citation patterns of *CL*, investigating which journals are cited in *CL* and which cite *CL* in turn. We find that *CL* shows higher preponderance to cite cognitive literature than other linguistic journals do, but that it is rarely being cited outside *CL*. Both analyses suggest that there are important opportunities to strengthen the links between research in *CL* and the rest of cognitive science. We end on a discussion of the theoretical importance of maintaining strong links across the cognitive sciences, and offer some initial ideas how cognitive linguists can close this gap and increase their role in the cognitive science community.

4.1 Introduction

Since its early beginnings the 1970s, cognitive linguists have stressed that language cannot be adequately studied in isolation from the cognitive processes that drive it. Their own work on language is inspired by and resonates with work that is done in other fields, especially cognitive psychology and to a lesser extent, cognitive neuroscience. Though cognitive linguistics could be regarded as its own branch of cognitive science and linguistics, some researchers in cognitive linguistics have emphasized that it is not a clearly defined field, but rather a body of scholarly work based on a set of core tenets (Geeraerts, 2006). One of these tenets, first formulated by Lakoff (1990) in the inaugural issue of *Cognitive Linguistics*, is the *cognitive commitment*:

> The cognitive commitment is a commitment to make one's account of human language accord with what is generally known about the mind and the brain, from other disciplines as well as our own. [...] The cognitive commitment forces one to be responsive to a wide variety of empirical results from a number of disciplines. (Lakoff, 1990, p. 40)

Thus, the study of language should reflect the state of knowledge about other cognitive processes and how they relate to language. To achieve that, the commitment calls for the integration of theories in other brain and cognitive sciences, such as neuroscience and psychology, making it arguably more interdisciplinary than general linguistics (see also Gibbs, 1996; Croft & Cruse, 2004; Evans, Bergen, & Zinken, 2007). It has also been argued that cognitive linguistics is unique from other areas of linguistics in that it is very much concerned with the general cognitive mechanisms that motivate linguistic form (Geeraerts, 2006). It also "seeks to discover the actual contents of human cognition" (Gibbs, 1996, p. 49). Cognitive linguistics has been stated to be "one of the principal branches" of cognitive science (Sinha, 2007).

In short, within cognitive linguistics itself, the following claims serve to set it apart from other areas of linguistics:

1. Cognitive linguistics studies general cognitive processes behind language use, while other forms of linguistics tend to neglect or deemphasize these cognitive factors.
2. Cognitive linguistics is an interdisciplinary field that integrates current findings from other cognitive and brain sciences, and links them to language.

In this paper, we examine these claims from a data-driven perspective, using the methods of scientometrics, the quantitative study of science

communication (Leydesdorff & Milojević, 2015). Beyond the theoretical notions on which these general claims are based, we ask whether they can be detected in the work of cognitive linguistics through quantitative study of publication patterns. To do this, we make a few assumptions. We assume that the journal *Cognitive Linguistics* (*CL*), as a flagship journal, is representative of the field of cognitive linguistics. We thus analyze papers published in this journal beginning with its inception in 1990. More specifically, we look at the following hypotheses arising from the two theoretical claims above:

1. The contents of *Cognitive Linguistics* are more similar to that of journals in cognitive science than those in general linguistics, as *CL* should span topics closely allied to other fields in cognitive science.
2. The articles in *Cognitive Linguistics* should cite work from cognitive science, and be more similar in citation behavior to articles in cognitive science journals than those in general linguistics, reflecting its closer link to the cognitive and brain sciences.

For the purpose of the current study, we have collected data from *Cognitive Linguistics* as well as the following journals: *Linguistic Inquiry* (generative linguistics), *Language* and *Lingua* (general linguistics), *Psychological Science* (psychology), *Cognitive Psychology* (psychology), *Cognitive Science* (cognitive science), *Cognition* (cognitive science), and *Metaphor & Symbol* (figurative language). To ground our analysis, we selected two journals from a more distant field, biology (*Ecology* and *Plant Physiology*). For each journal, we collected the abstract text and cited works (where available) beginning with 1990 (the first issue of *Cognitive Linguistics*).

In what follows, we take two analytic approaches to test these hypotheses: Using abstract content, we compare the similarities of journals with each other. Naturally, journals within the same field are expected to overlap in content more than journals in different fields. For example, neither *Ecology* nor *Plant Physiology* is expected to overlap in topics with any of the other journals. Second, we analyze the cited works in these journals. *CL* is expected to cite the cognitive science and psychology journals to a considerable degree, while also relying on some linguistics journals. In addition, we wanted to explore whether work in *CL* is taken up by psychologists and cognitive scientists, measured by the extent to which these journals cite *CL*.

In the General Discussion, we identify general "meta-theoretical" implications of our findings. The role of cognitive linguistics in cognitive science should be a central one, given both the importance of language as an aspect

of the human cognitive system, and the importance of cognitive linguistics as an interdisciplinary approach to language. Our results suggest that fruitful integration of *CL* with wider cognitive science has not yet been achieved, but there has been some progress. We will render two recommendations for *CL* to widen its impact. The first is to further integrate literature from other journals into articles in *CL*. The second is to widen the terminological and conceptual bridges between *CL* and recent cognitive research.

4.2 Content analysis

4.2.1 Data and methodology

In this section, we analyze the content of *CL* papers in comparison to other journals. We assumed that abstracts of papers offer an easily accessible approximation of the overall content of the paper, a strategy that has shown success in previous research (see Griffiths & Steyvers, 2004). We applied Latent Dirichlet Allocation (LDA) to the abstract text (see Chapter 2, Blei et al., 2003). LDA is a probabilistic Bayesian model that assigns topics (a collection of words) to documents (here, the abstracts). From these documents, the method infers a set of topics, each of them of a collection of words that are likely to occur within that topic. In this way, any original document can be represented as a distribution over topics that represent the gist of what the document is about. For example, a paper in *CL* could consists of the topic of motion and the topic of verbs, whereas a paper in *Plant Physiology* might talk about the topic of DNA and the topic of growth. After these topics have been extracted, they can be analyzed in more detail by looking at the most probable terms occurring within that topic. LDA has been successfully used previously to study scientific topics (Griffiths & Steyvers, 2004; Blei & Lafferty, 2007; Hall et al., 2008; Yau, Porter, Newman, & Suominen, 2014; De Battisti et al., 2015), and in this paper, we closely follow the procedure as described by Griffiths and Steyvers (2004), implemented in R (see Grün & Hornik, 2011; Ponweiser, 2012).

We followed standard procedure for pre-processing the text data. All words occurring in fewer than 5 abstracts were removed, as were words with fewer than 3 letters, and highly frequent words (stopwords). Removing these words removes the bias by low-frequency words, as well as removes distributionally pervasive short words, which are less likely to provide structure to the inferred set of topics. Each word was stemmed using the Snowball stemmer to account for basic morphological differences in word forms, such as number and tense (Porter, 2001). For example, the both the words “cognitive” and “cognition”

Table 4.1: Number of abstracts per journal.

Journal	Number of abstracts
Cognitive Linguistics	465
Lingua	1310
Language	340
Linguistic Inquiry	524
Metaphor & Symbol	364
Cognition	2319
Cognitive Science	977
Cognitive Psychology	439
Psychological Science	2283
Plant Physiology	3734
Ecology	727

become "cogni" after the stemming algorithm has been applied. This allows researchers to better compare words based on their stem and meaning, rather than their morphological features. Table 4.1 shows the number of abstracts per journal. While the numbers differ across journals, each journal is represented by several hundred abstracts, providing a basis for LDA to extract topics.

In LDA, the best number of topics k is not known a-priori. To find the best number of topics, and thus the model that best fits our data, the LDA algorithm was run with a different number of topics, $k = \{50, 100, 200, 300, 400, 500, 600, 1000\}$. Model fit was determined by the log-likelihood, and the model with the highest log-likelihood fit was chosen (see Figure 4.1). The best number of topics was determined to be 300 ($logLik = -6298439$). The following analyses are based on this topic model.

4.2.2 Topics in *Cognitive Linguistics*

As mentioned above, each document is assigned a topic distribution. For example, a document might mainly talk about two topics such as semantics and historical change, while another document covers mainly syntax and historical change. Topics in turn are represented by word distribution, that is, we can access the most probably terms/words for each topic. It is important to note that topics as such are not given titles – they are simply represented by a number (e.g. topic 1, topic 5, ..., topic k). These numbers are not in any particular order. By inspecting the terms, researchers can assign their own titles to the topics. While most topics can be intuitively interpreted by looking at their associated terms, some are less coherent than others.

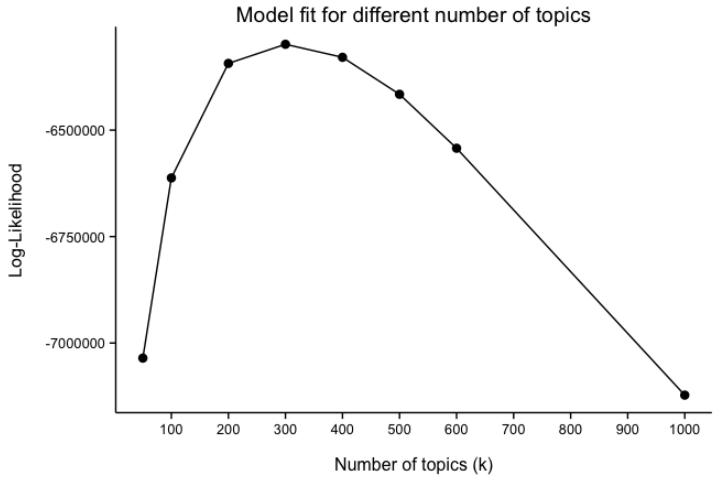

Figure 4.1: Model fit by number of topics. Highest log-likelihood represents best fit, and was reached at 300 topics.

After selecting the model of best fit, we looked at the 10 most common topics associated with papers in *CL*. This allowed us to do a manual inspection of how well the model fits the data, and to compare the topic distribution to other journals. We expected an overlap of these topics with the other linguistics journals, as well as with the cognitive science and psychology journals. Table 4.2 shows the five most probable terms for each of the 10 most probable topics. The most probable topic for *CL*, topic 164, relates to construction grammar (Goldberg, 1995; Goldberg, 2006). The second topic covers more general linguistic terms, relating to verbs, while the third topic identifies figurative language and metaphors (see Lakoff & Johnson, 1980). Topic 292 covers typology and language acquisition, and topic 20, grammar and syntax. Topic 74 covers word meaning, and topic 53 conceptual and abstract approaches. The remaining topics cover linguistic topics such as crosslinguistic studies (36), motion language (151) and discourse analysis (253). From this quick inspection of the most probable topics, we conclude that the LDA model offers an intuitive extraction of the scientific topics in *CL*.

How often do these *CL* topics occur in the other journals? We should expect some overlap with the general linguistics journals, as well as the cognitive science and psychology journals. In Figure 4.2, we plotted the probability each topic occurs in each of the journals, where a darker color represents a higher probability. Rows represent journals, and columns represent extracted LDA topics. We normalized the probabilities so that columns sum 1, such that the

Table 4.2: The ten most common topics in *CL* and their five most probable terms.

Topic 164	Topic 162	Topic 167	Topic 292	Topic 20
construct	verb	metaphor	languag	syntact
german	argument	conceptu	linguist	semant
analysi	semant	articl	acquisit	structur
english	lexic	text	typolog	grammat
paper	predic	convent	univ	grammar

Topic 74	Topic 53	Topic 36	Topic 157	Topic 253
sen	concept	language	event	speaker
metonymi	conceptu	english	encod	utter
cognit	abstract	speaker	path	communic
polysemi	cogni	bilingu	motion	convert
term	embodi	differ	manner	refer

probability of each tile is proportional to how often a certain topic is covered in that journal with respect to the other journals. This provides a visualization of the extent to which a topic tends to occur in one or more journals. dThe probability matrix shows that *CL* has greater overlap with linguistics journals than it does with cognitive science journals. Even the generative journal *Linguistic Inquiry* shows a higher similarity to *CL* than *Cognitive Science*. As expected, the biology journals show very little overlap. The analysis of the most probable topics shows that *CL* tends to cover linguistic topics rather than topics in cognitive science journals. One of the main interests of cognitive linguistics, figurative language, is monopolized by another journal, *Metaphor & Symbol*.

4.2.3 Correlation of topics

In the previous section, only the ten most common topics in *CL* were analyzed. Here, we look at all 300 topics and calculate the correlation across journals, allowing us to to compare journals based on their overall content: A high correlation expresses high similarity between journals, a low correlation dissimilarity. The correlation matrix is plotted below in Fig. 4.3 (red = high correlation, blue = low correlation). The matrix is ordered using hierarchical clustering (Ward algorithm), and the clusters are illustrated by black rectangles around them. Again, we see that *CL* exhibits a stronger similarity to linguistics journals than to cognitive and brain sciences journals. The clustering algorithm places *CL* firmly in the linguistics cluster.

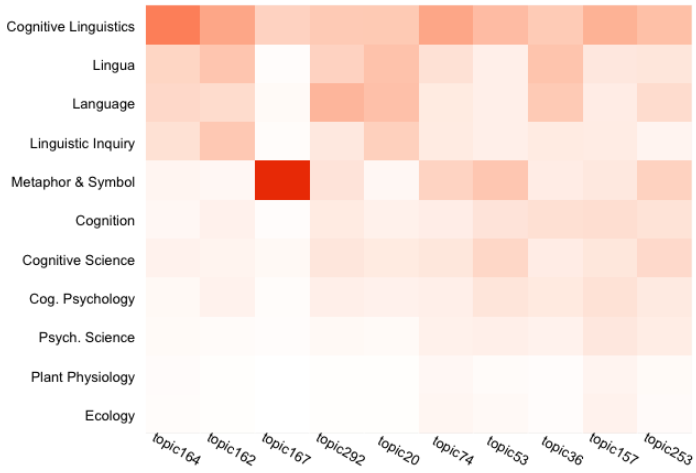

Figure 4.2: Comparison of the most popular topics in *CL* to other journals. The darker a tile, the higher the probability of that topic for the respective journal.

The linguistics journals are all highly correlated to each other, although *Linguistic Inquiry* is not as correlated to *CL* as the other journals, it is a small but significant relationship. When compared to the cognitive journals, we see that *CL* and *Language* correlate with the content of the cognitive journals to the same degree (none of their correlation values differ by more than 0.04). However, *Psychological Science* is negatively correlated to all linguistic journals, meaning the content is significantly different. Both *Lingua* and *Linguistic Inquiry* are not as highly correlated to the cognitive science journals, which shows us that that variation exists within linguistics journals.

4.2.4 Diagnostic topics

An alternative way to look at the topic distribution is to find the most diagnostic topic. The most diagnostic topic is the topic that can be considered unique to one journal, that is, the probability of it occurring in other journals is relatively low. This allows us to exclude general topics such as about scientific methods from our analysis, as well as more general linguistic topic that might wash out the relationship between journals covering topics. Following Griffiths and Steyvers (2004), we calculated the most diagnostic topic for each journal by dividing the probability of a given topic in one journal by the probability of the same topic in all other journals. The highest ratio thus denotes the most representative topic. Table 4.3 shows the most probable terms for these diagnostic topics.

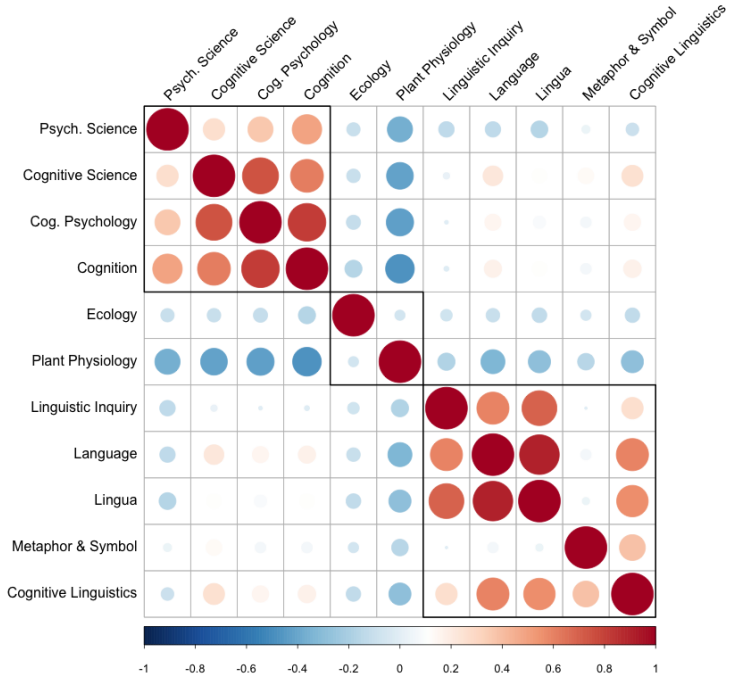

Figure 4.3: Correlation matrix showing similarity between journals.

We then plotted a similarity matrix to see how often these diagnostic topics occur in the other journals (Figure 4.4). Akin to the findings in Griffiths and Steyvers (2004), we should see clusters of similar journals. Cognitive science journals are expected to cluster together, as are the general linguistic ones. Based on the results of the previous section, we hypothesize that *Cognitive Linguistics* is more likely to fall into the cluster of general linguistics rather than cognitive science. We see a strong cluster around cognitive science and psychology journals, and a second cluster around linguistics, including CL. The diagnostic topics for the linguistic journals show a low probability of occurrence in other journals, including cognitive science, while cognitive science topics seem to be slightly more represented in the linguistics journals. However, *CL* does not stand apart from the other linguistics journals. Interestingly, the most representative topic for *CL* is the topic covering construction grammar, which is also represented in the other linguistic journals to some degree. Again, *Metaphor & Symbol*, as a highly specialized journal, has a very diagnostic topic which hardly occurs in other journals, not even *CL*.

Table 4.3: Most diagnostic, representative topic for each journal and their five most probable terms.

CL Topic 164	*Lingua* Topic 259	*Language* Topic 292	*Ling. Inq.* Topic 80	*Met. & Sym.* Topic 167
construct	claus	languag	deriv	metaphor
german	complement	linguist	movement	conceptu
analysi	head	acquisit	argu	articl
english	emb	typolog	articl	text
paper	predic	univ	analysi	convent

Cognition Topic 299	*Cog. Sci.* Topic 260	*Cog. Psych.* Topic 98	*Psych. Science* Topic 113	*Plant Phys.* Topic 201	*Ecology* Topic 72
infant	learn	memori	risk	mutant	speci
monthold	statist	retriev	depress	arabidopsi	communiti
experi	learner	recal	health	phenotyp	diver
month	acquisit	test	symptom	gene	trait
test	acquir	encod	childhod	plant	rich

4.2.5 Discussion

In both analyses, *CL* behaved more similarly to general linguistics journals than to journals in cognitive science. In addition, one of its most common topics, Topic 167, which is about figurative language, is dominated by *Metaphor & Symbol*. In our correlation analysis, we see some variation within the linguistics journals: Both *CL* and *Language* are correlated to cognitive journals, while the other two linguistics journals are not as much. So at least in view of the quantitative analysis of abstracts, hypothesis 1 – that *CL* is as or more related to cognitive science than to other linguistic journals – is not supported from these findings. Neither is *CL* more similar to these cognitive journals than are other linguistics journals: The correlation values of Language are very similar to those of *CL*, suggesting that *Language* is "as cognitive" as *CL*. More importantly, these correlation values to the cognitive journals are not nearly as high as those within the linguistics journals, meaning all linguistics journals, including *CL*, cannot be considered cognitive with respect to the technical terminology that is used. As a result, the clustering algorithm, based on the correlation values, places *CL* in the linguistics cluster, underlining that *CL* is not statistically unique among other linguistic journals.

One important caveat of this analysis is that our LDA analysis cannot extract deep conceptual relationships among the individual goals, methods and conclusions of each journal article. While the same topics are covered in linguistics

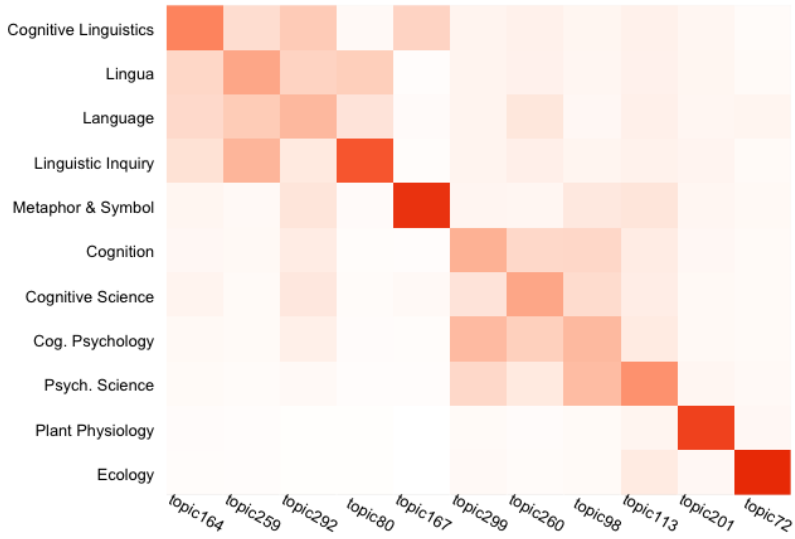

Figure 4.4: Most diagnostic topic for each journal, and their probability in the other journals. The redder a tile, the higher the probability.

journals and *CL*, this does not mean that they approach these topics under the same framework or treat the topics in the same way, meaning the conclusions can vary to a large degree. However, the data clearly show that topics in cognitive journals do not strongly characterize the content of *CL*, lending support to the alternative hypothesis that *CL* may have drifted from its cognitive roots. In fact, out of the 10 most probable topics, only one specifically involves cognition (topic 53), which concerns abstract and conceptual grounding. This result is underlined that when removing the word cognitive from the input text, the correlation values decreased across the board, suggesting that the correlation is partly driven by simply using the word 'cognitive' (as in *Cognitive Linguistics*) rather than actual content.

One additional caveat for this result is that, since *CL* is aiming to explain the same concepts as the other linguistics journals (albeit in a different way), the prominence of linguistics terms in *CL* might wash out its relationship to the other disciplines. However, as our analysis takes all content into account and is not restrained to linguistic topics, this is unlikely. If cognitive terms were present in *Cognitive Linguistics*, they would have been picked up by the LDA algorithm. In the next section, we evaluate the citations in *CL* and the other journals. It is

possible that, while *CL* does not highlight cognition terms in its abstracts, it still draws inspiration from work covering cognition.

4.3 Citation analysis

4.3.1 Data and methodology

Citations have been used previously to measure interdisciplinarity (Kreuzman, 2001; Porter, Cohen, David Roessner, & Perreault, 2007; Leydesdorff & Goldstone, 2014), and they are seen as an important contributor to the interconnected basis of scientific knowledge (Hyland, 1999). Within cognitive science, Goldstone and Leydesdorff (2006) looked at the import and export of the journal *Cognitive Science*: Which journals are cited by *Cognitive Science*, and in turn, which ones cite *Cognitive Science*? Their analysis revealed that *Cognitive Science* mainly imports from neuroscience and psychology, while in turn, it is mostly cited by cognitive psychology journals and computer science journals. Furthermore, *Cognitive Science* plays an import role in connecting scientific fields that otherwise are poorly connected, for example, by linking education and developmental psychology.

What might the import and export network look like for *CL*? Here, we mainly look at the import of *CL* – what journals *CL* cites – and only briefly consider its export. According to the tenets underlying cognitive linguists, we expect a high number of citations to journals in cognitive science and psychology. In particular, *CL* should exhibit a citation pattern that is as or more similar to these cognitive journals than to linguistics journals. We collected a list of cited works for the papers in our database. Unfortunately, citation data was not available for some of these papers, and not at all for the journal *Metaphor & Symbol*. In addition, citation data was only available starting from 2005. The following analyses are thus based on fewer individual papers per journal (see Table 4.4).

4.3.2 Import: Cited works

In this section, we evaluate the claim that cognitive linguistics integrates work from cognitive science, which should be reflected in the work *CL* cites – its import. To analyze this pattern, we constructed a journal-journal matrix per journal, counting the number of times a journal cites other journals (or more generally, other works, including books). Based on this document-term matrix we calculated a similarity matrix using cosine similarity. In this matrix, similarity between two journals is expressed by a value between -1 and 1, where 1 denotes high similarity

Table 4.4: Number of abstracts per journal with citation information.

Journal	Number of papers	Number of references
Cognitive Linguistics	208	10302
Cognition	2543	121090
Cognitive Psychology	784	41420
Cognitive Science	1162	60560
Ecology	701	35981
Linguistic Inquiry	203	11988
Language	203	16309
Lingua	1649	80740
Plant Physiology	3820	232818
Psychological Science	2266	67308

and -1 highest dissimilarity. Figure 4.5 shows this matrix. Cognitive Linguistics is more similar to the general linguistic journals in its citation behavior than it is to cognitive science journals. Again, we see a similar pattern to the clusters found in the content analysis: linguistics journals cluster together, including *CL*, and cognitive science and psychology journals form another cluster. This analysis is based on all cited works within a journal, no matter if it is a journal, book, or any other kind. In other words, the journals and books cited by articles in *CL* are, in general, more similar to the patterns of citation in other linguistics journals.

Whereas *CL* shows a high similarity to all linguistics journals, including *Linguistic Inquiry*, of all linguistics journals, it also has the highest similarity to the cognitive and brain sciences journals. What is more, all other linguistic journals show a negative correlation to these journals. Although the correlation values are very small (and much lower than the correlation to other linguistics journals), this shows that out of these linguistic journals, *CL* relies the most on works from cognitive sciences, and seems to import at least some theoretical or empirical work from the sources that undergird the cognitive and brain sciences. And although *CL* shows a similarity in citation patter to the cognitive journals, the content analysis showed that this does not translate into similarity on a content level. One possible explanation is that, while cognitive work is cited and applied to the study of language, very little is then mentioned when it comes to the consequences for general cognitive mechanisms and processes.

In a more detailed analysis, we constrained the cited works purely to the eight journals in our database, neglecting other citations. This selection does not include all works from linguistics and cognitive science, but it can be considered a reasonable representative sample, especially within the cognitive sciences. If

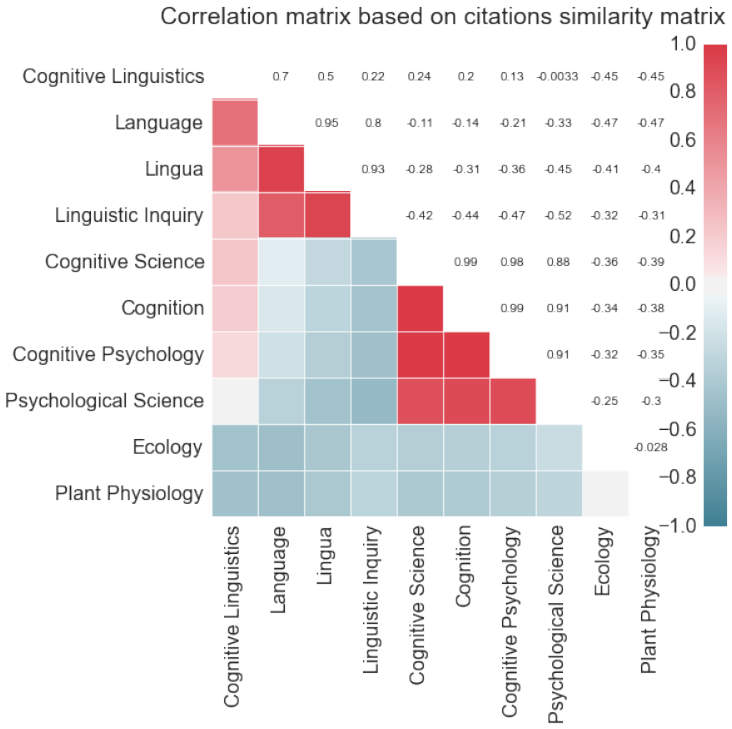

Figure 4.5: Similarity of cited works between journals.

CL is strongly interdisciplinary, it would be likely to cite the four cognitive science journals more than the other linguistics journals do. For each journal, we calculated the percentage of citations to the cognitive journals (*Cognition*, *Cognitive Science*, *Cognitive Psychology*, Psychological Science) and the linguistics journals (*Lingua*, *Language*, *Linguistic Inquiry*). The percentages are displayed in Table 4.5. Note that *CL* does not cite the four cognitive science and psychology journals any more than Lingua or *Language* does, but both journals rely on the other linguistics journals more than *CL* does. It does appear that *CL* is the more balanced among the language-specific journals; it relies less on the other three linguistics journals, even though it still shows a preponderance of citations to the linguistics journals over the cognitive journals. This suggests that a large number of cited works in *CL* are not part of the chosen eight journals, an important caveat of this follow-up analysis.

Based on the pure number of citations, we plotted a matrix visualizing the number of times a journal is cited by another journal (Figure 4.6). The dark diagonal represents self-citations of one journal citing itself. Clusters are

Table 4.5: Percentage of citations coming from cognitive and linguistics journals, ranked by percentage of cognitive journals citations.

Journal	% CogSci	% Linguistics
Cognition	16.2	5.45
Cognitive Science	9.78	2.81
Cognitive Psychology	6.64	1.25
Psychological Science	6.48	0.1
Lingua	0.94	35.13
Language	0.43	7.22
Cognitive Linguistics	0.43	2.17
Linguistic Inquiry	0.1	10.8
Ecology	0	0
Plant Physiology	0	0

present for cognitive sciences, and to a slightly lesser degree, the linguistics journals. This confirms that the results from the previous analyses that *CL* does not cite cognitive journals predominantly; however, it does seem to rely less on the linguistic journals, suggesting that other sources not present in our analyses are cited frequently.

4.3.3 Export

The goal of an interdisciplinary journal is not only to import work from other fields, but also to influence them in turn by exporting ideas and theories. Many measures exist to assess the impact of a journal, but we restrict ourselves to analyzing how often *CL* is cited by the journals in our database to detect whether works from cognitive linguistics are in turn taken up by psychologists and cognitive scientists. Advances made in the study of linguistics and theories developed by cognitive linguistics should, of course, contribute to the knowledge base of other cognitive scientists.

Table 4.6 shows the number of times a journal cites *CL* (since 2005). The data reveal that *CL* accounts for nearly 60% of all *CL* citations within these eight journals. Within the other journals, *CL* accounts for less than 1% of citations within the respective journal. However, the degree to which *CL* itself is not unusual, with citations to *CL* only accounting to 3%. As a comparison, self-citations account for 3.39% in *Cognitive Science* and 5.56% in *Cognition*. Neither of the linguistics nor cognitive science and psychology journals cite cognitive linguistics to a large degree, suggesting that *CL* does not export to a large degree and is at risk of becoming, at least by this analysis, somewhat isolated in its impact. On a positive

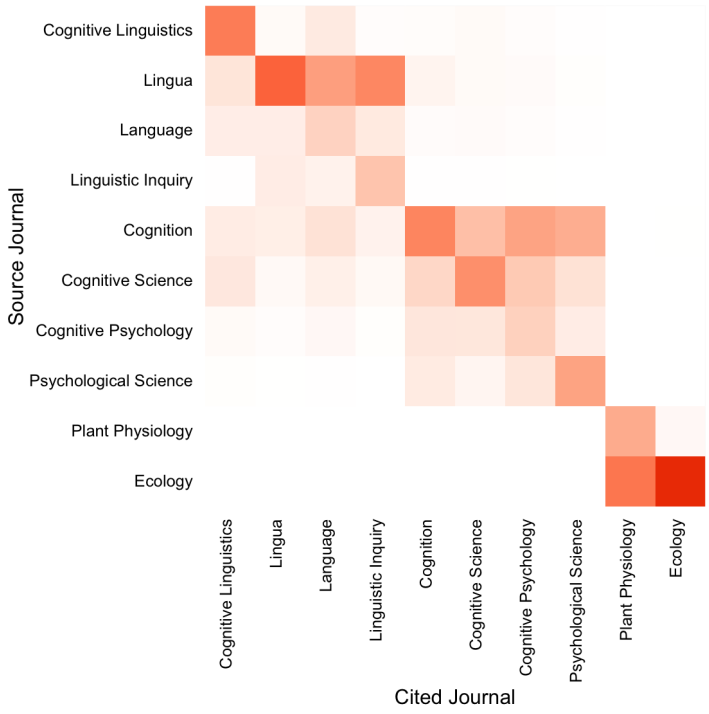

Figure 4.6: Citation count matrix. Each tile represents the number of times a journal cites another journal. The darker the tile, the higher the citation count.

note, we see that *Cognitive Science* and *Cognition* cite *CL* more than some of the linguistic journals, suggesting that at least some work from *CL* is taken up in the broader cognitive science community.

4.3.4 Discussion

Both the import and export of *CL* reveal a journal that is not importing or exporting as widely as one would hope, given its bold tenets. On balance, its journal articles still rely more on linguistics work than on work in cognitive science, while rarely exporting to either cognitive or linguistic journals. On balance though, *CL* cites more outside the journals in our database than the other linguistics journals do, implying that its citation base is wider than the citations in the selected linguistics journals. However, in all our cluster analyses, *CL* falls squarely in the domain of linguistics, both in content and citation.

We chose the journals carefully as representative of their respective fields. However, they do not cover the full spectrum of cognitive science, neuroscience

Table 4.6: Export of *CL*: Number of *CL* papers cited by other journals. The third column denotes the percentage of refernences in a journal that are to *CL*, the fourth column the percentage of all *CL* citations in the analyzed journals.

CL cited by	Count	Within-Journal (%)	Of-*CL* (%)
Cognitive Linguistics	312	3.03	59.32
Lingua	61	0.09	11.60
Cognitive Science	53	0.10	10.08
Cognition	44	0.04	8.37
Language	39	0.24	7.41
Cognitive Psychology	12	0.04	2.28
Psychological Science	3	0.00	0.57
Linguistic Inquiry	2	0.02	0.38

and psychology. It is therefore possible that other journals cited *CL* belong to these fields, however, the fact that two of the biggest journals of cognitive science are not represented more strongly gives cause for concern. The citation analysis suggests that *CL* shows a higher similarity in citation patters to cognitive journals than the other linguistic journals, however, as our content analysis showed, this does not reflect in content similarity.

4.4 General discussion

CL, on balance, is more similar to other linguistics journals than cognitive journals. Given the topics that *CL* covers, this may come as no surprise. After all, it is a framework within linguistics, and it would be even more surprising to find few links to other linguistics journals. Furthermore, *Cognitive Science* and *Cognition* are not the only journals in which research in this more general domain can be found. Recently, there have been both arguments for and against the claim that *Cognitive Science* is an interdisciplinary journal (Goldstone & Leydesdorff, 2006; Bergmann et al., 2016). Some have argued that it has become dominated by psychology (Gentner, 2010), and some fields, such as anthropology, have become largely separated and independent from cognitive science (see Beller, Bender, & Medin, 2012; Keen, 2014). That different topics are represented in *Cognitive Science* and *Cognitive Linguistics* could just as easily be attributed to researchers in *Cognitive Science*, as well as the low number of citations of *CL*.

Even though our data analysis does not support hypothesis 1 (that the content of *CL* is as or more related to cognitive science than to other linguistic journals), hypothesis 2 (that CL relies on more inspiration from cognitive sciences)

gained some support: *CL* showed a slightly higher percentage of citations to cognitive science journals than other linguistic journals, except for *Language*. The fact that this does not relate in higher overlap in content overlap could mean that, while works from cognitive science serve as an inspiration and are applied to language, it does not translate into detectable discussion about general cognitive theories and processes. This notion is supported by the export pattern of *CL*, which suggests papers in *CL* are mostly cited by other papers in *CL*. In general, in both our analyses *CL* behaves very similarly to Language, a journal covering general linguistics. This suggests that *CL* does not have a special place among linguistics journals in terms of how cognitive it is, but rather, fits in with the broader field of linguistics that has taken up interdisciplinary work. Taken together, the content and citation analysis indicate a separation of *CL* from wider, more general cognitive science journals. Regardless of the causal basis of this apparent separation – if it is worrisome at all – it would seem helpful both to cognitive linguistics and cognitive science more broadly to encourage deeper interconnections and overlap in concepts and topics. This requires both participation and engagement and, importantly, to have the broader cognitive science community listen. What might help is investment in citing and discussing work in other journals more extensively, reaching back to the root of *Cognitive Linguistics*. This investment could bring great returns, as cognitive linguistics has much to contribute to our understanding of both language and wider cognition. Psycholinguists, communication scientists, and other researchers interested in language could be helped by cognitive linguists, who might support wider cognitive science research by putting it on a sounder linguistic foundation.

We perceive two ways in which *CL* may widen its impact and draw new bridges to other fields. The first obvious recommendation is to integrate more recent cognitive literature in the theories and empirical work developing in *CL*. As we show above, the preponderance of *CL* citation is to linguistics journals, although it also cites some cognitive literature. Wider citation patterns may be achieved by seeking inspiration from cognitive science research beyond *CL*. It may also derive from identifying emerging issues in cognitive science journals for which *CL* ideas would have much to say. Such patterns of citation and conceptual inspiration would modulate the scientometric measures we have shared here. A second, more important recommendation is to renew a shared terminological or conceptual foundation with the cognitive sciences. One way of achieving this is to develop empirical and computational intersections. Notice that the topic space for *CL* is still squarely in the traditional methodological realms of linguistics itself: the terms and topics of a strictly qualitative linguistic

enterprise. Obvious terminological bridges can be achieved by taking up new methodologies in these other cognitive fields, including empirical and computational techniques. Of course there are clear examples of both in the *CL* literature (e.g. Gonzales-Marquez, Mittelberg, Coulson, & Spivey, 2007; Regier, 1996). But the analysis here suggests that the dividends paid from this methodological integration may still be great. Furthermore, both of these approaches would likely help to increase the interest of cognitive scientists in the work of cognitive linguistics; at the moment, the low export of *CL* shows that cognitive linguistics is a rather isolated enterprise.

We wish to briefly consider an obvious criticism of what we have presented: Does our quantitative analysis even recommend worrying about these issues? Can topic analyses and citation patterns even tap deeply enough into *CL* to warrant the attention of readers and authors of *CL*? An obvious cosmetic riposte to such a criticism is that it is in the best interest of *CL* to appear integrative according to these text-based statistical analyses. The future of document search and retrieval, disciplinary intersections and quantification (e.g. Rinia, van Leeuwen, Bruins, van Vuren, & van Raan, 2002) – strongly recognized by funding agencies – recommend such a cosmetic strategy: Putting the interdisciplinarity front and center by sharing terms and literatures. Another theoretical concern derives from cognitive linguistics itself. The power of words to weave particular conceptual intersections, particular conceptual framing, and so on, is well known from cognitive linguistics research itself (Lakoff, 2004). Seeking a robust shared space of terminology may best frame the future of *CL* to offer new advances in understanding language and cognition that can be understood and integrated by other fields, as well.

In conclusion, it is worth sharing the final thoughts of another scientometric analysis, of Higgins and Dyschkant (2014), who made similar arguments about interdisciplinarity in philosophy. In the following quote, "philosophers" is to be exchanged with "cognitive linguists", and "nonphilosophers" with "non-cognitive linguists":

> Philosophers should communicate and collaborate with nonphilosophers, attend nonphilosophy conferences, exchange ideas with other academics, and coauthor works with experts in these fields. We believe that such actions will lead to significant practical and intellectual benefits for philosophy and academics more generally.

CHAPTER 5

Scientific structure across time

This chapter is an updated version of the following manuscript:
Bergmann, T., Dale, R. & Bhat, H. S. (manuscript). *Comparing patterns of change in science and the humanities*.

5.1 Introduction

Philosophy of science has long debated how to define scientific progress and how this progress happens (for an introduction, see Losee, 2004). While different views of progress exist, all imply that a later step is better than a previous step, not just simply change (Niiniluoto, 1980). What is seen as "better", however, differs among philosophers and is one of the most controversial topic within philosophy of science. Not only are there different theories on what constitutes progress (Bird, 2007), there is also debate about how this progress — regardless of definition — comes about. For a long time, scientific progress was seen as continuous and steady, rather than "revolutionary" (Wray, 2006). Under this view, knowledge or truth (depending on definition of progress) is cumulatively added to the existing body of knowledge. This view has been criticized as "naive and oversimplified" (Niiniluoto, 1980, p. 429), as progress is not always linear. Kuhn's influential work (1962) also rejected this view. Instead, scientific change is attributed to shifts in *paradigms*, which completely replace previous views and methods. These paradigm shifts do not happen gradually, but suddenly and abrupt. The introduction of an irrational factor into science was met with criticism by various philosophers of science (see Lakatos & Musgrave, 1970), yet Kuhn's work proved to be seminal.

In this chapter, we approach this debate with a quantitative angle. Can we measure scientific progress by relying on computational methods and the

output of scientists themselves? Because progress inherently includes a qualitative element – a later step is seen as better than an earlier step – we focus on analyzing *change*, rather than progress. Can we use scientometric methods to measure the type of change in scientific literature? In order to answer this question, we study how automatically extracted topics and themes of published articles in the fields of philosophy and biology change in popularity over time. Philosophy and biology were chosen because they differ in many ways, and thus serve as a good base for comparison. For example, while biology relies on quantitative analyses based on a variety of data, philosophy is more qualitative and subjective of nature. Papers in biology also tend to be co-authored by multiple authors, relying on large-scale, cross-departmental collaboration. Philosophy, on the other hands, produces longer papers authored usually by one sole author. Such structural differences are likely to lead to differences in how the field behaves in their scientific output, and thus how this output changes over time.

Furthermore, both within and outside of philosophy, there has been some debate whether philosophy as a field makes progress. Within philosophy, these positive changes have often been defined as conclusively answering "big questions" – questions that define the field and that the majority of the researchers in a field are interested in. This is congruent with the Kuhnian view of science as problem solving. After answering such a big question, usually a new big question appears – thus the set of big questions changes over time. However, in the view of some philosophers, these big questions do not get answered and stay the same (Nielsen, 1987). Dietrich (2011, p. 322) argues that philosophy "is the exactly the same today as it was 3000 years ago", and if Aristotle were to sit in on a philosophy lecture today, he would fully understand the subject matter, and even be able to participate. In contrast, such advances have been made in biology (and other natural sciences) that he would not even understand concepts that are now basic. Dietrich's view is corroborated by a survey of leading philosophers on a set of thirty big questions such as "Is there a god?" and "Do we have free will?". Answers on these questions were heavily divided with very few having an answer that was not controversial (Bourget & Chalmers, 2014).

On the other hand, there is little doubt or debate whether biology has made progress, or changed over time. Impactful discoveries such as the structure of DNA (Watson & Crick, 1953) have opened new avenues of research, and similarly, old theories such as Lamarckism have initially abated in their popularity, only to be revived again later in modified fashion (Jablonka & Lamb, 2005; Moore, 2015). Other advances have led to the emergence of new fields, such as synthetic biology, the combination of biology and engineering (Oldham, Hall, & Burton,

2012). Large scale, international collaborations such as the Human Genome Project and advances in technology have made it possible to study biological systems in a novel way. In fact, there is little discussion within the field of biology and society at large whether biology progresses.

Because problems in biology are more "answerable" than in philosophy, the fields might also change in different ways. If questions in philosophy are hardly ever conclusively answered, we might expect their popularity to fluctuate over time. On the other hand, in biology, problems can be solved and new problems emerge. Thus, change should happen more linearly and not fluctuate over time as much. The pattern of change might differ between the two fields.

Using abstracts of scientific articles as our data set, we will attempt to quantitatively answer the following questions:

1. Can we model scientific change using abstracts from scientific papers?
2. Does biology undergo more change than philosophy?
3. Does the pattern of change in biology differ from the pattern in philosophy?

In the next section, we will describe our data set in more detail, and then move onto the methodological explanation.

5.2 Dataset

The dataset consists of a sample of papers from biological and philosophical journals, provided by JSTOR. In total, 8314 philosophical papers and 15596 biological papers were collected, including their abstracts and year of publication. The papers span the time range from 1980–2013, as unfortunately, earlier years were not represented well enough to warrant inclusion in this analysis. Each year has at least 30 papers published during that year in each of the two fields. Fig. 5.1 shows the sample size of papers per year and per field. The distribution of papers over time is not uniform, but increases linearly over time, with a small drop in 2013. However, the sample size per year was large enough to not bias any analysis.

The samples generated for each field included a variety of representative journals. The philosophy sample was generated from 66 different journals, and the biology sample from 174 journals. The five most frequent journals in each sample are shown in Table 5.1. While some journals occur more often than others, the linear pattern observed in Fig. 5.1 holds for the individual items, meaning that there is not a sudden increase in one journal at a particular time. This makes it possible to compare different years to each other, irrespective of journal.

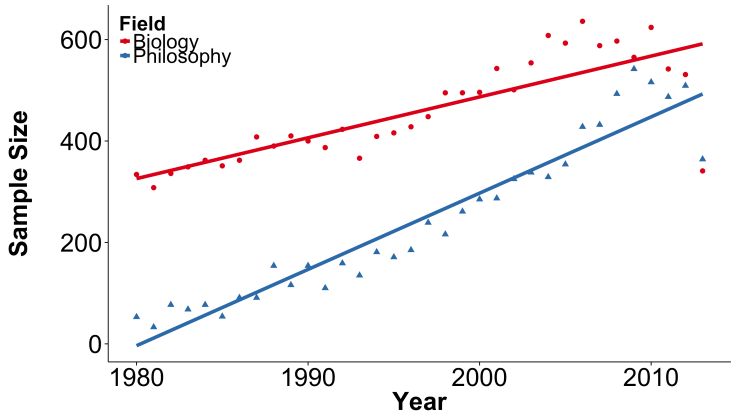

Figure 5.1: The number of papers per field over time.

The text of abstracts, the key variable in our analysis using topic modeling, was present for all the papers we selected. Following common procedure in natural language processing, each abstract was tokenized and stemmed. Common words (*stopwords*) such as *this* and *it* were also removed, as were short words below three characters. Fig. 5.2 shows an example of an abstract in its original and processed form.

Original

Looking at Thomas Kuhn's work from a cognitive science perspective helps to articulate and to legitimize, to some degree, his rejection of traditional views of concepts, categorization, theory structure, and rule–based problem solving. Whereas my colleagues focus on the later Kuhn of the MIT years, I study the early Kuhn as an anticipation of case–based reasoning and schema theory. These recent developments in cognitive psychology and artificial intelligence may point toward a more computational version of Kuhn's ideas, but they also expose ambiguities in his work, notably in his understanding of exemplars.

Stemmed

thoma kuhn work cognit scienc perspect help articul legitim degre reject tradit view concept categor theori structur rulebas problem solv colleagu focus kuhn mit year studi earli kuhn anticip casebas reason schema theori recent develop cognit psycholog artifici intellig point comput version kuhn idea expos ambigu work notabl understand exemplar

Figure 5.2: Original (left) and processed (right) article. The processed article strips the abstract of all unnecessary text that does not contribute to the overall meaning of the abstract.

Table 5.1: Five most frequent journals in each sample.

	Journal	Number of papers
1	*Journal of Business Ethics*	1529
2	*Synthese*	696
3	*Philosophy of Science*	584
4	*The Journal of Symbolic Logic*	541
5	*Studia Logica*	410

	Journal	Number of papers
1	*Plant Physiology*	934
2	*The Journal of Cell Biology*	750
3	*Oecologia*	515
4	*Ecology*	446
5	*New Phytologist*	380

5.3 Topic modeling as a scientometric tool

A swath of variables have been used in scientometric analyses, including author-provided keywords, co-author collaborations and both the import and export of citations. Each measurement has its advantages and disadvantages, and captures a different angle of a publication. Capturing the semantic and scientific content of an article is no easy task, as one single variable does not accurately reflect the whole content of an article. Currently, most research in scientometrics uses abstracts as the key data when analyzing content, as they provides more detail than simple keywords and are available for most publications. Unlike keywords, abstracts are also entirely written by the authors, while journals occasionally force the authors to select keywords from a provided list. Although abstracts provide the best approximation of content, they are hard to analyze. They cannot be summed up in one single word or number, and contain various information. For instance, abstracts often describe both the method and the results of a study. Hand-coding such information is not feasible, especially with large corpora. Instead, advances in natural language processing have made it possible to extract latent variables from textual data.

The most recent suite of algorithms are *topic models*, which discovers latent themes and topics within texts (for overviews, see Griffiths, Steyvers, & Tenenbaum, 2007; Blei, 2012). The most basic and original form of topic modeling is *Latent Dirichlet Allocation* (LDA) (Chapter 2, Blei et al., 2003). The basic underlying idea is that each document comprises multiple topics, and each word

in a document is assigned a topic. Each topic is thus a distribution over terms, that is, words that are likely to occur in that topic. An example would be articles in a newspaper. Articles in newspaper cover a range of different topics, such as economics and sports. LDA assumes that each article is biased to talk about certain topics, rather than the whole range of topics. For example, an article in the sports section might talk about championships and winning, while an article in the economy section will rarely talk about these topics. In turn, such topics are also biased to include different terms. A topic about championships will be biased to include terms such as ring or cup, compared to words such as income or budget, which are more likely to occur in an economics topic. The same principle holds for scientific articles. For example, a topic about DNA might include terms such as *gene* or *cell*. Each document in turn is modeled as a distribution over topics: Some topics are more probable to occur in the document, and others are very unlikely, and thus have low occurrence. While a paper in biology has a high probability of including a DNA topic, a paper in philosophy does not.

LDA uses probabilistic algorithms to infer these posterior distributions of topics and terms, and to allocate each document a topic distribution, and each topic a term distribution. When applying topic modeling to documents, each document will be represented by a distribution over topics. More information about the topics can be learned by looking at its associated term distribution, and topics can such be summed up in a coherent manner. After the distributions have been inferred, each document is now represented as a distribution over topics θ_d, and each topic can be inspected in more detail by looking at its associated term distribution ϕ_z. This matrix representation allows us to perform computations on documents, such as comparing similarities between documents and calculating popularity of topics over all documents.

Topic modeling has found wide applications in the field of digital humanities, for instance, in analyzing historical newspapers (Yang, Torget, & Mihalcea, 2011) and classical scholarship (Mimno, 2012). It is also used widely in the field of political science (Lucas et al., 2015). Within scientometrics, topic modeling has been applied to variety of problems. For example, it has been used to detect similarities between different fields (Griffiths & Steyvers, 2004; Yau et al., 2014) and measure interdisciplinarity (Nichols, 2014). Several case studies investigate the change of topics over time in specific fields, such as computer science (Hall et al., 2008; Anderson, McFarland, & Jurafsky, 2012), cognitive science (Cohen Priva & Austerweil, 2015) and statistics (De Battisti et al., 2015). Here, we will use a similar approach and model the change of topics over time in different fields, instead of looking at isolated topics.

Table 5.2: Five most frequent topics and their terms in biology sample.

Topic 49	Topic 2	Topic 71	Topic 72	Topic 64
process	gene	soil	effect	model
system	express	concentr	increas	data
function	mutant	nutrient	respons	estim
studi	regul	product	result	method
ecolog	arabidopsi	organ	affect	predict
import	plant	nitrogen	reduc	base
interact	signal	carbon	level	sampl
structur	respons	rate	experi	measur
role	transcript	biomass	control	test
understand	protein	increas	competit	analysi

5.4 Applying LDA

For each of our two sample, we fit a LDA model with $T = 75$ topics. The number of topics has to be chosen a-priori, and previous studies in the field of scientometrics suggest 50-100 as a good number of topics to produce topics that are neither too broad or narrow (see Hall et al., 2008). Fitting two separate models instead of one combined models makes it possible to study both samples independently. In a combined model, the higher number of abstracts in biology would exert too much influence on the topic model compared to the philosophy sample.

After running the topic models, we assessed the most frequent topics in each field to make sure the algorithm produced an adequate represenation of the abstracts. Table 5.2 and Table 5.3 show the five most frequent terms and their associated topics for biology and philosophy respectively.

As each document is now represented as a probability distribution over 75 topics and the year of publication is known for each document, we can trace the popularity of a topic over time by looking at its probability for each document at a given year. However, as the magnitude of popularly cannot be assumed to be linearly dependent on the year of publication, non-linear methods need to be used. The next section will introduce natural cubic splines and explain how they can be used to model the change in probability values over time.

Table 5.3: Five most frequent topics and their terms in philosophy sample.

Topic 28	Topic 2	Topic 25	Topic 22	Topic 72
ethic	studi	corpor	set	logic
busi	result	respons	show	modal
issu	behavior	social	texmath	semant
practic	signific	firm	degre	complet
manag	find	csr	prove	proof
develop	influenc	stakehold	theorem	predic
articl	attitud	perform	exist	oper
decisionmak	research	compani	infinit	formula
execut	perceiv	manag	cardin	rule
field	effect	invest	result	system

5.5 Fitting natural cubic splines

5.5.1 Natural splines

Clearly, the relationship between time and the popularity of a topic is not linear and can not be modeled simply by using a linear regression. Such a model might wash out local minima and maxima and not represent the change in time accurately. One way to combat this is to add polynomials predictors to the linear regression, however, as this is applied globally, this sometimes has unintended consequences in the border regions of the predictor space. Instead, one can divide the predictor space X into several regions and fit separate low-degree polynomials over these regions. Such regions are bound by *knots*. Regression splines improve these models even further by constraining the coefficients in such a way that the fitted curve is continuous, i.e. there is no jump at knots. Each separately fitted polynomial connects to the curve in the region before and after. This is done by ensuring that both the first and second derivative of the piecewise polynomials at knot points are continuous. To produce more stable estimates at the boundary regions of X, a further constraint can be introduced: The function has to be linear at the boundary. Such splines are called *natural splines*, and will be used here (for introductions on these models, see Hastie, Tibshirani, & Friedman, 2009; James, Witten, Hastie, & Tibshirani, 2013).

More formally, a polynomial spline of degree D with K knots at locations $\xi_1, \ldots, \xi_K$ is defined by the following function:

$$y = \beta_0 + \sum_{d=1}^{D} \beta_d x^d + \sum_{k=1}^{K} b_k (x - \xi_k)_+^D \tag{5.1}$$

where

- $(x - \xi_k)_+^D = 0$ when $x < \xi_k$ (to the left of the knot)
- $(x - \xi_k)_+^D = (x - \xi_k)^D$ when $x \geq \xi_k$ (to the right of the knot)

This means that the predictor matrix X_{mat} for a spline of degree D with K knots is as follows (assuming that X is one-dimensional):

$$X_{mat} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^D & (x_1 - \xi_1)_+^D & \dots & (x_1 - \xi_K)_+^D \\ 1 & x_2 & x_2^2 & \dots & x_2^D & (x_n - \xi_1)_+^D & \dots & (x_n - \xi_K)_+^D \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^D & (x_n - \xi_1)_+^D & \dots & (x_n - \xi_K)_+^D \end{bmatrix}$$

The second column is simply the original value of the predictor x. We then add columns of polynomials of x up to the specified degree D. Lastly, predictors are added according to the indication function $(x - \xi_k)_+^D$, again up to the specified degree D. An estimate of y can then be simply calculated using linear regression, $\hat{y} = \beta X_{mat}$.

For a natural cubic spline, we perform a similar transformation, but instead, the basis function $N(x)$ for a natural spline with degree D is defined as:

$$\begin{aligned} N_1(x) &= 1 \\ N_2(x) &= x \\ N_{k+2}(x) &= d_k(x) - d_{K-1}(x) \end{aligned} \tag{5.2}$$

where

$$d_k(x) = \frac{(x - \xi_k)_+^D - (x - \xi_K)_+^D}{\xi_K - \xi_k} \tag{5.3}$$

and $(x - \xi_k)_+^D$ is the same indicator function as above. The predictor matrix can then be represented as follows:

$$X_{ns} = \begin{bmatrix} 1 & x_1 & d_1(x_1) - d_{K-1}(x_1) & \dots & d_K(x_1) - d_{K-1}(x_1) \\ 1 & x_2 & d_1(x_2) - d_{K-1}(x_2) & \dots & d_K(x_2) - d_{K-1}(x_2) \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & d_1(x_n) - d_{K-1}(x_n) & \dots & d_K(x_n) - d_{K-1}(x_n) \end{bmatrix}$$

After obtaining X_{ns}, we can again fit a linear regression with the matrix as the predictor, $\hat{y} = \beta X_{ns}$.

To illustrate the different fits of a cubic spline and a natural cubic spline and to justify our choice of natural cubic splines, models with $D = 3$ and $K = 2$ were fitted to the same artificially generated data. The knot locations are equispaced between the minimum and maximum of the predictor values. The fits are plotted in Figure 5.3, along with a linear regression fit. The knot locations are denoted by the dashed vertical lines. While the linear regression fit does not take into account the local peak in the center of the predictor values, the cubic spline predicts higher values in that regions. However, it behaves quite extremely at the boundary regions, with the predicted values sharply decreasing at the lower end of the predictors and sharply increasing at the higher end. This is especially a problem for out-of-sample predictions in the boundary regions. The natural cubic splines fits its curve more smoothly at the boundaries and yet takes into account the rise in values in the center.

While the natural spline in the example above gives the best and most natural fit, the real element of interest in this study is whether there is a change in predicted values at any given point in time. That is, at a given value for x, is our value predicted to change or stay the same? In simple linear regression, we can assess this change by examining the slope. If the slope is sufficiently different from zero, it indicates either positive or negative change over time. However, with splines, there is not one single coefficient that can be examined. Instead, we can look at the *gradients* of the fitted line at a given point in time - if there is no change, the line will be straight and the gradient zero.

To find the derivate of a natural spline, the derivative of $N(x)$, $N'(x)$ needs to be calculated. We can then use the derivative to calculate gradients given an x-value. As both $N_1(x)$ and $N_2(x)$ are constants, they are zero when derived, which leaves the derivate of $N_{k+2}(x)$ (see Equation 5.2). For this, the derivative of $d_k(x)$ (Equation 5.3) needs to be calculated as follows:

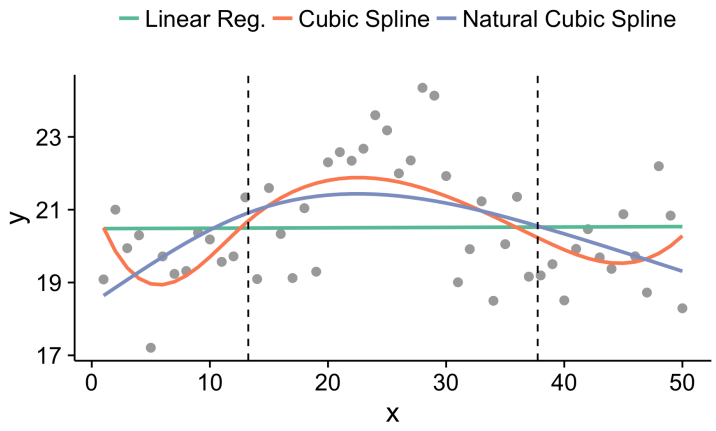

Figure 5.3: Linear regression, cubic spline, and natural cubic spline fitted to the same data. The linear regression washes out the local peak in the middle of the predictor. The cubic spline behaves erratically in the boundary region, but takes into account the local peak. The natural cubic spline combines the local peak with smoother fits in the boundary regions. The dashed vertical lines show the locations of the knots.

$$d'_k(x) = \frac{D(x-\xi_k)_+^{(D-1)} - D(x-\xi_K)_+^{(D-1)}}{\xi_K - \xi_k} \tag{5.4}$$

Figure 5.4 shows the gradient of the natural spline (Panel B) next to the natural spline fit (Panel A). The dotted grey line is drawn at $y = 0$, denoting a zero gradient. When the gradient is above said line, the change is positive. When the gradient is below zero, the change is negative. When the gradient intersects the line, the direction of change flips: In this example, it goes from positive to negative.

5.5.2 Applying to topic distributions

For our data, we chose to set the degree D to 3, a common choice that allows flexibility while not overfitting to the sample data. In the model, we only had one predictor, the year of the publication, and one outcome variable, the probability of the topic in that document. A separate model was fit for each topic. As our outcome variable is a probability, we applied the logit function to its values:

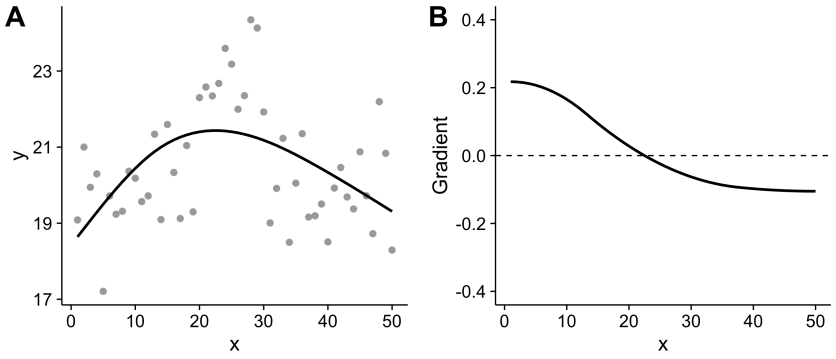

Figure 5.4: Panel A shows the natural spline fit. Panel B shows the gradient for the natural spline.

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

Subsequently, the predicted values from the natural spline are transformed using the inverse logit (sigmoid) function:

$$S(p) = \frac{1}{1+e^{-p}}$$

This ensure that the predicted values are in the interval $[0, 1]$, and thus only return possible probabilities.

To select the number of knots, we ran 10-fold randomized cross-validation on a number of knots for each topics. The number of knots varied from 1 to 6, and for each number, it was assumed the knots would be equispaced from each other. The reasoning behind this is that it is impossible to know a-priori which periods in time would be particularly susceptible to change, and indeed, there is no reason to assume that one particular period should be. Furthermore, the maximum of six knots was chosen as we assumed that significant change in science cannot be accurately represented from year to year, and thus we assume a minimum period of five years for change to be represented in journal articles. As our years range from 1980–2013 and cover 34 years, this results in a maximum number of six knots that are at least five years apart. After running the cross-validation, the minimum number of knots that resulted in the best mean squared error (*MSE*) was chosen for each topic.

To obtain confidence intervals for both the spline and the gradients, 2000 bootstrap iteration were run. A 95% confidence interval was then calculated from the 2000 samples of predicted values and gradients respectively using the

2.5% and 97.5% quantiles. Figure 5.5 shows the obtained confidence intervals for both the natural spline fit (left) and the gradients (right) of one sample topic. If the confidence intervals of the gradient do not contain zero, we can say that the gradient is positively or negatively significantly different from zero. In the example, the gradient is significantly different from zero most of the time, with a few stable points in time when changing direction. The gradients can thus be converted into a binary variable that denotes whether a topic is changing (positively or negatively) at a given time. In the example, the topic probability is decreasing 60.7% of the time, increasing 17.5% and stable at the remaining 12.8%.

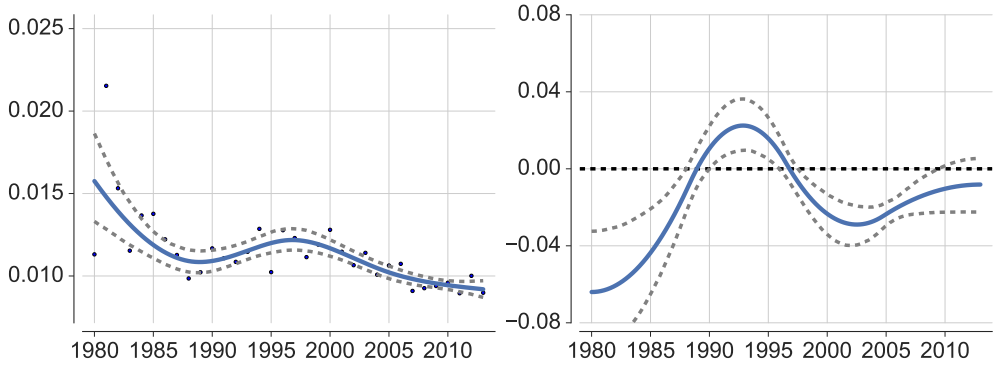

Figure 5.5: Left panel shows the natural cubic spline fit (blue line) for a sample topic (topic 72 in philosophy data set) after transformation. The dashed lines represent the confidence interval. The right panel shows the gradient in blue, with the confidence interval in the dashed lines. The horizontal dashed line denotes a zero gradient. Gradient is significantly different from zero if confidence intervals do not overlap with this line.

By applying this method to all topics across the two data sets, we can extract the following information:

1. The proportion of overall changes per topic and per field.
2. The proportion of negative and positive changes per topic and per field.
3. The years during which each field exhibits change, and the years during which they are more stable.

This allows us to quantify the patterns of change in both philosophy and biology. Do both fields exhibit the same proportion of overall change? Is this change positive or negative? Do the topics behave linearly in their popularity, or do they wildly fluctuate in popularity? Using the methodology explained above, the next section attempts to answers these questions.

5.6 Analysis and results

5.6.1 Can we model scientific change through splines?

A simple, yet effective test to check whether the models with only year as a predictor capture at least a degree of variability is to shuffle the years that are associated with each publication. After re-running the spline models on the shuffled year data, no trends should be detected. To test this hypothesis, both the number of knots selected in the cross-validation and the final gradients were examined.

A simpler structure in the shuffled year data should result in fewer selected knots: The less change there is, the fewer knots are necessary to obtain a good fit to the data. Indeed, a paired one-sided t-test showed significant differences both for philosophy ($t(74) = 3.59, p = 0.0003$) and biology ($t(74) = 2.47, p = 0.0079$). This indicates that the topic changes over shuffled years are indeed less complex than in the original data.

A similar intuition is true for the gradients: If the topic changes are less complex, the gradients in the shuffled years spline model should be clustered around zero, whereas in the original models, potential trends should be represented by non-zero gradients (see also below). Figure 5.6 shows histograms of the absolute values of the gradients for both models in both fields. Absolute values are shown because the direction of change is irrelevant in this case. Gradients in the shuffled years model are more biased towards zero, while the gradients in the original model have a higher magnitude. To test these difference for statistical significance, paired one-sided t-tests were run on the absolute values of the gradients. Significant differences were found both for philosophy ($t(74999) = 115.13, p = 0.00$) and biology ($t(74999) = 271.47, p = 0.00$).

This simple analysis shows that topics indeed vary significantly over years, and the model of choice, natural cubic splines, effectively capture the trends and changes of topics over time.

5.6.2 Does biology change more than philosophy?

Using the methodology explained above, we can look at several different features of scientific change. Our first hypothesis was that biology, as a field, changes more on average than philosophy. As we have identified the degree of change at any given point in our time range, we can analyze the proportion of non-zero changes. Table 5.4 shows these proportions. Within biology, nearly 80% of the gradients are significantly different from zero, while only 58% of philosophy gradients are

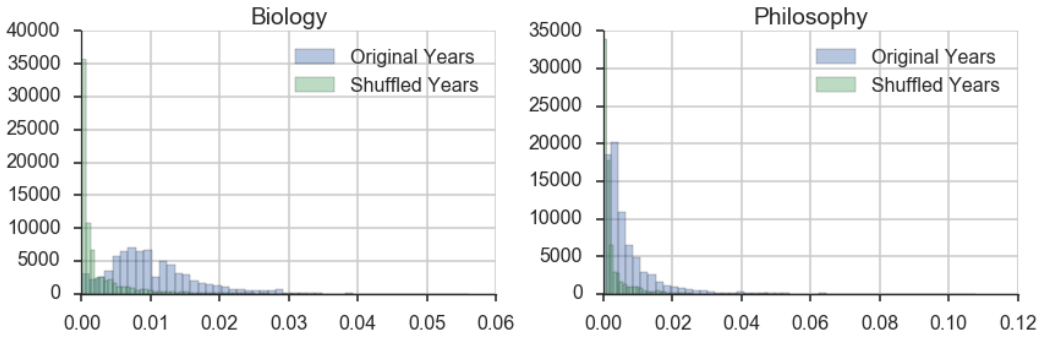

Figure 5.6: Comparison of absolute values of gradients in the original model and the shuffled years model. When years are shuffled, gradients are more clustered around zero, indicating that no trends are detected. This difference is more noticeable in biology.

Table 5.4: Percentage of gradients in each field that are non-zero (95% CI).

Field	Increasing	Decreasing	Total	SD(Total)
Biology	8.85%	70.49%	79.34%	33.52%
Philosophy	12.12%	45.80%	57.92%	23.81%

non-zero. In both fields, these changes are largely driven by a downward change, indicating that new topics hardly ever increase at a rapid pace, instead, they are more likely to meander upwards. The higher proportion of change in biology is significantly different from philosophy ($t(148) = 4.48, p < 0.0001$). The first hypothesis is thus confirmed: Scientific topics in biology change more often than topics in philosophy.

However, one drawback of this analysis is that it does not take the magnitude of the change into effect. A non-zero change can be significant, but small. While the above analysis reveals that biology changes *more often*, it does not tell us *by how much*. By summing the absolute values of all non-zero values, the total magnitude of changes per topic can be calculated. There was a significant difference in the total magnitude for biology topics (M=9.56, SD=3.51) and philosophy topics (M=5.4, SD=5.06); $t(148) = 5.8, p < 0.0001$). Figure 5.7 shows the distribution of magnitude values for both fields, and clearly, biology shows higher values, although some fields in philosophy also undergo large changes.

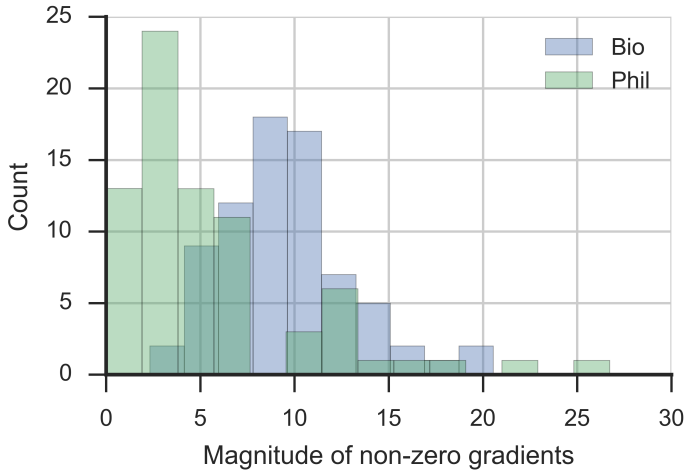

Figure 5.7: Distribution of absolute values (magnitude) of non-zero gradients. Biology gradients have a higher value than philosophy, on average.

Table 5.5: Number of topics in each field that exhibit certain patterns of change.

Field	Only Increasing	Only Decreasing	Both	No change
Biology	6	59	10	0
Philosophy	6	36	27	8

5.6.3 Are the patterns of change different?

Another hypothesis is that changes in biology might be more linear than in philosophy, where topic popularity fluctuates over time. In biology, it is more likely that problems get "solved", that is, an answer if found. Scientists can then move onto the next problem, or topic. In philosophy, this is not as easy, as definite proof for a lot of problems is not obtainable. To investigate these questions, we looked at the direction of change more closely.

A topic that fluctuates exhibits both increasing *and* decreasing changes. Out of the 75 topics in each field, 27 in philosophy both increase and decrease at different times. Only 10 topics in biology behave in such a way. Conversely, eight topics in philosophy do not change at all, while all topics in biology undergo change (Table 5.5).

In a similar vein, if a topic goes from increasing change to decreasing change, the gradient changes its sign from positive to negative. If a topic fluctuates in popularity, we can expect a lot of sign changes. If a topic stays stable, or only

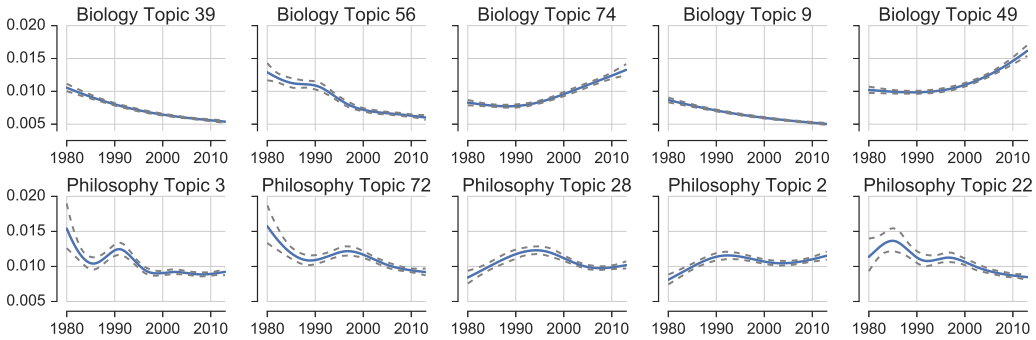

Figure 5.8: The five topics per field that undergo the most change. Topics in biology undergo a more linear change, while topics in philosophy have more local peaks and troughs.

changes in one direction, the number of sign changes should be close to zero. The previous results indicate that we should expect more sign changes in philosophy, and indeed, this is the case. There was a significant difference in the number of sign changes for biology topics (M=0.89, SD=1.37) and philosophy topics (M=1.48, SD=1.55); $t(148) = -2.44$, $p = 0.016$.

Using results from the analysis above, we can select topic that undergo the most change in each field. For scientists and scholars within this field, such an analysis can reveal latent structures and changes within their field. For example, in our data-set, such a selection reveals that out of the five most changing topics in biology, three continuously decrease, and two exhibit a surge in popularity after 1990 (see Figure 5.8 and Table 5.6). Both of these, topic 74 and 49, can be tied to a recent interest in understanding impact on ecology and preserving and conversing nature.

On the other hand, the five topics in philosophy that change the most exhibit more fluctuation, and cannot be easily summarized as *increasing* or *decreasing*. Each topic has local *bumps* in popularity, after each the values decrease again. Two of the topics, topic 72 and 22, seem to generally decrease in popularity, and are tied to more formal analysis of philosophy (see Table 5.7). These difference in changes within these five topics are congruent with the overall analysis above, which suggests that philosophy is subject to more "random" fluctuations in topic popularity.

Table 5.6: Associated terms with the topics in biology that undergo the most change.

Topic 39	Topic 56	Topic 74	Topic 9	Topic 49
membran	activ	manag	cultur	process
cell	acid	conserv	medium	system
transport	text	land	line	function
plasma	enzym	research	embryo	studi
vesicl	concentr	develop	develop	ecolog
surfac	inhibit	protect	transfer	import
calcium	extract	human	produc	interact
protein	metabol	assess	growth	structur
cytoplasm	accumul	area	media	role
fraction	level	natur	vitro	understand

Table 5.7: Associated terms with the topics in philosophy that undergo the most change.

Topic 3	Topic 72	Topic 28	Topic 2	Topic 22
scienc	logic	ethic	studi	set
scientif	modal	busi	result	show
philosophi	semant	issu	behavior	texmath
philosoph	complet	practic	signific	degre
epistemolog	proof	manag	find	prove
empir	predic	develop	influenc	theorem
scientist	oper	articl	attitud	exist
methodolog	formula	decisionmak	research	infinit
histor	rule	execut	perceiv	cardin
histori	system	field	effect	result

5.7 Discussion

The quantitative analysis has brought up several points. First, topic modeling has proved itself to be a good technique to extract the themes of philosophy and biology in our sample. The manual inspection of topics found clear and cohesive topics which are easily interpretable. Second, we can track the popularity of changes using natural cubic splines, and analyze these changes. Both fields showed a high proportion of change within these topics, suggesting that neither field can be considered stagnant based on scientific publications. More in-depth analysis of these changes found the following results:

1. Topics in biology exhibit change at more points in time than topics in philosophy.
2. Non-zero gradients in biology have a bigger magnitude than in philosophy.
3. Topics in philosophy fluctuate more than in biology, while topic change in biology is more linear.
4. The methodology makes it possible to identify topics undergo a lot of change.

There are several possible why these results were obtained. There is a small chance that the decisions (such as selection of knots) made during the analysis produced the results. However, as several different avenues, in particular with respect to the number of knots, were explored, this is unlikely. Additionally, the large size of the data-set makes it improbable that these results are artifacts of only a few journals or papers. It is more likely that these differences in changes reflect a deeper, latent structure within the fields.

The first two results were in line with our prediction that biology undergoes more change than philosophy, and confirms theoretical discussions from philosophers themselves (see Chalmers, 2015). Philosophy makes fewer and less drastic changes than biology. As noted earlier in the paper, this is partly because biology is more driven by problems for which answers can be found, while many questions in philosophy cannot be conclusively answered. This gives more room for a potential re-emergence of a topic, whereas in biology, bringing up an already solved topic is usually pointless – unless there is new evidence or a re-analysis of some sorts. The results can also be explained with theories in philosophy of science. According to Kuhn (1962) and Laudan (1977), science is mainly concerned with solving specific problems and puzzles , which in our case seems to be an accurate representation of change in biology – as a puzzle is solved, it will no longer be actively worked on and new puzzles slowly emerge. Using our method, a puzzle thus roughly equates to a topic in our topic model. On the other hand, it might not be a good description of philosophy. Instead, other models of scientific progress such as the epistemic view (Bird, 2007), where progress is seen as accumulation of information and new methods, describe the changes in philosophy better.

Previous research in computer science has shown that topic changes can sometimes be tied directly to external events. Hall et al. (2008) showed that funding led to a sudden influx of topics that were funded by government grants, and their disappearance after the grants ran out. Similar external events can influence topic popularity in biology and philosophy. While it is less likely that grants have a huge impact on research topics in philosophy, other events such as

current social climate and social issues can affect what the "hot topics" are, and which topics are unpopular. Additionally, changes in topics could be driven by changes in technology. For example, the rise of Bayesian statistics can largely be attributed to the improvement in computational power, despite the underlying principles going back many centuries. While philosophy so far has not been transformed by this rise in computational methods, neighboring disciplines such as cognitive science and artificial intelligence have (Cronin, Shaw, & La Barre, 2003). Advances in those areas could provide new insight into the core areas of philosophy, or philosophy could take up those methods itself (although that seems unlikely). Changes in technology have also led to changes how science is published, for example, an increase in open access publication (Swan, 2007). Interaction with such computational systems can change over time, and thus, can prompt topical changes as well (see Gersh, Mckneely, & Remington, 2005; Preece et al., 1994; Carroll, 2006).

While we analyzed patterns of change more generally in this paper, researchers in specific fields can use such computational methods to understand their own field better. For example, the most changing topics (see Figure 5.8) can be further examined by experts in the field, and qualitative judgments be made about the developments. For example, not all change might be seen as progress, but rather as unwanted. A more local, detailed perspective on such patterns is thus possible and worthwhile to examine whether a field is moving into the "right" direction. It can also identify which topics are under-represented, and whether additional attention should be paid to encourage studies in such topics, e.g. increased funding and exposure.

Another key difference between philosophy and biology is the structure of research teams. There has also been continuing research and interest into the composition of teams and their performance (Goldstone, Roberts, & Gureckis, 2008; Woolley, Chabris, Pentland, Hashmi, & Malone, 2010). Group performance studies have emphasized importance of building appropriate combinations of skills and experiences on teams (Huber & Lewis, 2010), which poses interesting challenges to philosophy, due to the lack of of teamwork and interdisciplinarity on an author level. Content analyses such as presented here could provide an alternative measure to analyze relationships between academics, and how they group together. As Kreuzman (2001) has shown, acknowledgments are a common way in philosophy to attribute credit to other scholars, indicating some kind of group behavior. Combining both these measures could reveal clusters of academics working on similar topics and problems, and how these are tied to topical changes over time. Another possible extension to the current study

is to rely on citations. In the cited reference analysis of philosophy of science journals by Wray and Bornmann (2015), they found distinctive peak years of cited references, indicating that specific works in certain years are more impactful than others. A comparative analysis to biology would reveal whether philosophy relies more on older work. Natural science is more unlikely than philosophy to cite works from over hundred years ago, while philosophy relies more on classic works.

5.8 Conclusion

Our paper shows that scientific change can be modeled using computational tools. Using topic modeling and spline regression, extracted topics from scientific abstracts can be tracked over time. We have found that biology as a field changes more often, at a higher rate, and in a different pattern than philosophy. Nevertheless, philosophy is not a stagnant field, but topic popularity fluctuates over time, thus rejecting the view that philosophy does not change at all. Our findings has consequences for how we assess scientific progress, especially with regard to differences between the natural sciences and humanities. It can be combined with theories in philosophy of science to improve our current understanding of progress, and the debate between different views of scientific progress. Additionally, computational analysis like ours can serve as a tool for researchers to assess the state of their own field, as well as a way to quantify internal progress.

CHAPTER 6

General discussion

The three previous chapters have shown how topic modeling can be extended in fruitful ways to study different aspects of scientific systems and academic behavior. Chapter 3 combined topic model with social networks, studying the internal structure of a scientific conference. It identified strong clusters of topics, which equate to different sub-fields of the study of language evolution, and helped to identify areas where the multidisciplinarity of the main conference is lacking. With the creation of a new journal in the research area that aims to unify language evolution research, this showcases an important impact of such computational analyses: Based on the results, it is now possible for researchers in the field to emphasize areas that are part of the field of language evolution but currently under-represented.

Chapter 4 used topic modeling to analyze the structure and relationships between different scientific fields, evaluating claims from one community with quantitative data. The results indicate that – despite claims from researchers – the field of cognitive linguistics is not well integrated with the bigger field of cognitive science. Instead, it is nearly indistinguishable from other approaches in linguistics. Again, such a quantitative study can be used to re-evaluate the current state of a field – and whether that state is desirable. While some linguists do not find our results troubling, others have worked towards a bigger integration of cognitive and quantitative methods in cognitive linguistics (e.g. Gonzales-Marquez et al., 2007). Our analysis can help cognitive linguistics to restructure and reorient themselves, as well as foster relationships across discipline boundaries (if, at all, desired).

Lastly, Chapter 5 used topic modeling to track changes in scientific topics and discourse over time, combing LDA with cubic spline regression. Analyzing biology and philosophy as proxies for natural sciences and humanities, we

discovered different patterns of change over time within these fields. Changes in biology are more linear and either increase or decrease, while changes in philosophy have local maxima and minima, indicating that popularity of topics in philosophy is tied more to subjective phenomena such as current societal issues. In general, the case studies showed that topic modeling is not only a suitable tool for scientometrics, but that it can also be extended in a variety of ways to study different aspects of scientific organization. Additionally, it improves on earlier studies which often relied on keywords (e.g. Bentley, 2008) by using abstracts which are less constrained than variables such as keywords and subject codes.

These case studies show scientometric studies can gain insight into scientific communities, internally, externally, and temporally. While the conclusions of such case studies often still involve subjective, quantitative analysis by researchers with sufficient domain knowledge, quantitative data helps them to reach the right conclusions. For example, while Chapter 3 shows that the EvoLang community is dominated by one theoretical and methodological framework, whereas other areas such as archeology and genetics are not present at all. It is up to researchers within the community to change the status quo. However, armed with such results, other researchers in the community might be more susceptible to accepting attempts to change the field and include other areas of research. A similar case exists for the analysis of cognitive linguistics in Chapter 4.

The research presented in this dissertation also ties in to the question of how success in academia is defined and evaluated. Different approaches to quantify academic success have been proposed and are currently used. Often, measures such as citations for individuals (e.g. the *h*-index, Hirsch, 2005) and impact factor for journals are used to quantify the impact of publications. An even simpler approach is to simply look at the number of publications, which has led to the *publish or perish* paradigm. More nuanced views are taken up by other researchers, who identify which publications push the *cognitive boundaries* of fields – that is, use novel ideas and methods. Using the lexical diversity of publication titles, Milojević (2015) showed that while the number of scientific publications has risen exponentially, the number of cognitive domains only expanded linearly. Novel research does not always lead to higher citations, because not enough people might publish (yet) on that topic. On the other hand, if a large number of productive researchers publish a lot on the same, or a set of related topics, this will lead to more citations by design. For example, in Chapter 3, one cluster of research topic was shown to be dominant – by working on questions relevant to that cluster, you have a higher chance of being cited than working in a smaller area of comparative studies. However, this does not mean that the work is less relevant, or

less important, and as noted, researchers within the field have identified the lack of such studies as a weakness. Similarly, in cognitive linguistics (see Chapter 4), work emphasizing the link between cognitive science and linguistics might not be as highly cited within the cognitive linguistics community or the cognitive science community as work that is seen as more central to either community. However, such research might lead to novel results that can tell us more about the link between language and cognition. This tension between the goals of an individual and the goals of a community are further elaborated below. As the evaluation of researchers has a large impact on their subsequent career, finding the right measure is paramount, especially when citations are not an adequate measure.

While computational, quantitative analyses of science and academia are not necessarily new (e.g. Newman, 2001), a lot of studies are purely descriptive and neglect explanations behind their findings (e.g. Petersen, 2015). As evident from the previous chapters, scientific structures are complex: They involve many different agents in a hierarchical system, spanning different social and cultural environments and involve a temporal aspect. Each publication – our measure of interest in this dissertation – is the result of many decisions taken by the authors on the paper. These decisions are reflected in the quantitative analyses in the previous three chapters. Scientific publications do not emerge in a vacuum, they are actively pursued by academics. The authors decide what to work on, who to work with, and where to publish. But what factors influence their decisions, and how so? For example, while graduate students are commonly first authors, often the research topics and co-authors are heavily influenced by their advisor. A non-tenured professor is under different pressure than a tenured professor. Such environmental and social factors can heavily influence decision making. Researchers have studied these factors from different angles, including management studies, graph theory, cognitive science and others.

One interest of note has been group formation and composition. Here, *group* or *team* describe the list of collaborators who publish a paper together. The average team size of collaborators has steadily increased since World War II, along with a general increase in scientific output (Bornmann & Mutz, 2015). Most researchers believe that teams form to solve more complex problems because it requires interdisciplinarity, thus drawing researchers from different universities and fields, resulting in an increased team size across the board (Börner et al., 2010). But how are these teams formed, and how do they affect research topics and output? Chapter 3 has shown how group formation changes from conference to conference, and that the resulting social network of collaborators is tied to research topics. Research in graph theory has modeled how ideas spread through social

networks (Kempe, Kleinberg, & Tardos, 2003), and how groups in these networks are formed (Backstrom, Huttenlocher, Kleinberg, & Lan, 2006). Backstrom et al. (2006) find that the probability of an individual joining a community depends on how many friends (or, co-authors in the scientific realm) the individual has in that community. This relationship holds for online communities as well as for scientific conferences: A scientist is more likely to present at a conference if she has collaborated with many authors at the conference, and members in an online community are more likely to join a sub-community if they know many members in it. Additionally, if those friends are also friends with each other, they are even more likely to join. They also find that authors can be attracted to join a conference due to a "hot topic" of high popularity.

Other studies have investigated how cumulative team formation affects the topology of the resulting collaboration network (Guimera, Uzzi, Spiro, & Amaral, 2005). Over time, scientific collaboration networks transition from a collection of small disconnected clusters to one large cluster where the majority of nodes (around 70%) within the network make up one large component of connected nodes. This change from "disconnected school of thoughts" to an "invisible college" happens through iterative team formation, both through the addition of new team members and repeated collaborations with already existing team members. Similar results were obtained in the study of EvoLang authors in Chapter 3. In both studies, each node in the network has the same influence on other nodes, however, this is usually not the case in academia, where the lead author of a paper (usually) has higher control over both the paper content and team composition. This is taken into account in a newer study, which models team as both *core* and *extended* teams. Extended teams are core teams plus additional members, where those members can form their own core teams (e.g. if two principal investigators join forces along with their respective graduate students). Modeling teams in this manner, Milojević (2014) shows that the size of core teams rises only slowly, while extended teams expand more and more. The increase in the average size of teams is thus mainly driven by the emergence of "super-collaborations" of tens to hundreds of co-authors.

The rise of teams in science has also been analyzed with respect to their output. Do large-scale teams perform better than smaller teams or individuals? In general, most studies show that teams with more co-authors produce work that is cited more often than studies with smaller teams (Guimera et al., 2005; Milojević, 2014). More in-depth studies show more citations are associated with collaborations with distinct prior publication history (Bhat et al., 2015), long physical distance collaborations (Larivière, Haustein, & Börner, 2015) and

collaborations between different universities (Uzzi, Mukherjee, Stringer, & Jones, 2013). Guimera et al. (2005) also show that journals with a high impact factor consist of collaboration networks where the giant component includes the majority of authors. Even though large teams are now more frequent than before, small team collaborations still dominate the production of scientific articles in raw numbers, thus leading cumulatively to more impact (Milojević, 2014).

These studies show that team formation has substantially changed over time, and as they are tied directly to their citation count, we can also assume that team formation affects the content of papers. However, so far, no study has directly investigated this link. The three case studies in this dissertation have shown that scientific content depends on internal and external structure of a community and changes over time. Team composition could be one of the driving factors between all these results. As teams change over time, communication becomes more difficult and can lead to a decrease in performance (Huber & Lewis, 2010; Mao, Mason, Suri, & Watts, 2016). A change in communication can also lead to a change in research topics, especially when individuals in teams pursue different goals. Ideas and concepts live in an environment, or *habitat*, which influences how popular this concept is (Berger & Heath, 2005). Teams on a smaller scale and the social network they live in thus have an effect on which topics are selected to be worked on. For example, there might be pressure on graduate students to conform to expected behavior and culture, and not stray outside these bounds, limiting the range of topics they can work on. Additionally, every team member is part of a larger network of all researchers/authors, which can be seen as collectively storing available knowledge. The position of members in the network affects how they access information contained within the network and whether they have access to a diverse or limited set of resources (Guimera et al., 2005). Such factors could influence how cognitive linguistics has mainly stayed inside its conceptual boundaries (Chapter 4), or how fields differ in their change over time (Chapter 5).

The studies above show how the interaction with other team members, and a network of collaborators in general influences which topics are worked on. However, these studies lack a coherent theory on why these behaviors emerge in such settings. One overarching theory or framework to model these kinds of influences and behaviors is the notion of collective behavior and group cognition (Goldstone & Gureckis, 2009). The core tenet of collective behavior is that individuals rarely, if ever, act in total isolation from other individuals, and thus it is not sufficient to simply study individuals' cognition. It is not possible to predict the behavior of a group by simply studying the individuals on their own. The

behavior of one individual influences the behavior of others, and often, this can be indirect. For example, individuals often change the environment they interact with, which in turn influences other individuals interacting with this environment (*stigmergy*). A foot path over grass is often created by individuals repeatedly using it, where each individual slightly adds to the emerging path, which in turns make it more likely for other individuals to also use that path (Moussaid, Garnier, Theraulaz, & Helbing, 2009). Groups are thus often more successful at coming to a solution or creating a tangible output than single individuals (Theiner, Allen, & Goldstone, 2010). A common example of this is the *wisdom of the crowd*, where the mean estimate of a numeric quantity improves as more individuals give estimates. Social information – for example, information of the estimates of the other group members – can improve the overall accuracy of estimates, showing that collective behavior can be better than the sum of all individual guesses (Granovskiy, Gold, Sumpter, & Goldstone, 2015).

Collective behavior also informs us about how groups choose problems, and how they choose strategies to solve these problems. Studies from cognitive science have found that there is a systematic relationship between the difficulty of a problem and the optimal organization of a group trying to solve it (Goldstone, Wisdom, Robert, & Frey, 2013). Agent-based models have shown that agents quickly adopt the traits of their neighbors and aggregate into like-minded clusters of agents (Schelling, 1971; Axelrod, 1997). Another commonly observed pattern is that individuals converge to the solution of the group in estimating tasks (Granovskiy et al., 2015). In a study of music downloads, Salganik, Dodds, and Watts (2006) showed that individuals are more likely to download music that their peers have also downloaded, thus creating a snowball effect where few songs dominate the market. These examples show how group cognition can restrict the overall set of information and problems to a smaller subset available to the community. Depending on the task and problem, this can be either positive or negative, but for complex problems, this kind of *exploitation* of a previously found solution is usually negative (Mason & Watts, 2012; Goldstone et al., 2013).

Instead, a trade-off between *exploitation* and *exploration* is needed to find the best possible solution. Only exploration would result in underutilization of the current solution and to "re-inventing the wheel". Agent-based models show that densely clustered and locally connected agents in a network facilitate exploration, while the more exploitation is more suited to more globally connected agents. Well isolated groups can be helpful if different regions of a problem space need to be explored. Such groups cannot communicate with each other efficiently, and thus, cannot converge to one common (sub-optimal) solution. The more complex

a problem space is – for example, a problem that has multiple local maxima solution and one global maximum solution – the more exploration is needed. In some ways, these studies suggest that it is more beneficial if members in a network are weakly or not at all connected as they will not converge to a local maxima. However, it has been argued that agent based models do not model actual collective behavior well, as they are subject to constraints and assumptions that might not reflect the actual world accurately. In a large-scale experiment on Amazon Mechanical Turk, Mason and Watts (2012) connect different network topologies to problems with varying complexity. The experiment is essentially a foraging game where connected groups of individuals have to find resources. The resources are distributed according to the fitness function, and points are awarded for finding resources. This means that depending on the complexity of the fitness function and therefore the distribution of resources, exploitation or exploration is necessary to collect the maximum number of points. Participants in the experiment were aware of the location of their networked neighbors, and could take this information into account when choosing where to position themselves. In general, individuals connected through any kind of network perform better than independent individuals of the same size. Furthermore, centrally connected individuals in the network performed better (received a higher score) than more peripheral members. Depending on the complexity of the network, exploration can be harmful for individual members of the network, but helpful for the collective (increasing average group score). Exploration is often a high risk, high reward scenario where the individual member who is exploring will not always be rewarded. Exploration can however also lead to finding the global maximum, and subsequent exploitation by the collective. Members in local clusters tended to copy whatever their neighbors did, which decreased mean performance of the network, but led to an increase in scores for some of these individuals. Thus, there can be competition between an overall group goal and individuals goals. In contrast to other agent based models (see also Goldstone & Janssen, 2005), Mason and Watts (2012) found that participants in the experiment *always* performed better when in a well-connected network. If the global maximum (the best possible solution) was found, this information was spread more efficiently to the rest of the network, which in turn increased overall scores. Additionally, participants did not get stuck in local maxima, and still explored to a similar degree than in the low-connected networks. They conclude that the simple presence of a well-connected network that can communicate efficiently is not enough for a premature convergence on local maximum.

Many of the aforementioned results have direct bearing on academia and science. Collaborators form a connected social network, as seen in Chapter 3. The structure of this network has a direct influence on research topics, which can be explained by collective behavior: Research areas can be seen as complex problems with a complex fitness or solution landscape, and the way communication is transferred through the network can affect which solution is targeted by members in the network. While no consensus has been reached on which exact network topology is the most suited for scientific problems, it is clear that the relationship is rather complex. In the scientific world, it is made even more complex by the fact that communication does not simply happen through collaborations, but also through citations, reading papers, and verbal communication at conferences. Agent based models are not a suitable tool to accurately portray the complexities of scientific communication, and even large scale experiments are unlikely to portray the full complexity. Results from such studies are thus taken with a grain of salt, nevertheless, they can help shed light on certain aspects of academic organization and structure, and how they relate to dissemination of information.

The notion of exploration and exploration can also be applied directly to the selection of research areas and the number of citations publications receive. Scientists can either extend a current trend or an established theory in a field, or explore new problems and solutions. Exploration here is clearly tied to more risk: Exploring a new problem might not result in a publishable result and years of no scientific output. However, if a solution is successfully found, the impact of the resulting work is generally high. Academics who exploit established work productively, can expect a modest but steady return, while scientists can achieve high status by pursuing risky research (Foster, Rzhetsky, & Evans, 2015). While these achievements are on an individual level, based on the collective search studies we would expect risky research to be necessary for science to progress. In a large scale analysis of scientific papers in biochemistry, Foster et al. (2015) show that successful innovative papers receive a much higher citation count than conservative strategy of extending current work. A similar result is obtained in a citation study by Uzzi et al. (2013), in which papers with novel combinations of citations – e.g. drawing from different fields – have a higher impact than papers with conventional citations. However, even papers with novel combinations need a high base of conventional papers to receive citations, meaning that papers still need to be "anchored" in some way to established research.

Research areas are not the only resources available to researchers, as funding describes resources in a more traditional way. Researchers need to forage for these resources, just as they forage for research areas and topics.

Grants provided by various agencies are one of the main funding sources for scientists, and applying for grants requires a lot of time and usually does not end in successful funding (Bollen, Crandall, & Junk, 2014). Viewing funding as a foraging problem can help scientists to better allocate their resources, as well as help funding agencies decide who to fund. For example, features such as team composition and variety of cited references are related to the impact of scientific work, and can thus be used to predict future impact of grant proposals. Similarly, the research presented in Chapter 5 can be taken into account by both researchers and funding agencies to better predict which topics will be popular in the future.

Collective behavior and group cognition thus seems to be a fitting framework to study scientific behavior (see also Cronin, 2004). Academia is a complex adaptive system, in which agents both pursue individual and collective goals. Using theories grounded in cognitive science can reveal more about the behavior at the individual and group level, and combined with the computational tools used in this dissertations can shed light on both local (e.g. within a scientific field) and global (e.g. across all fields) issues in academia. Globally, it can help funding agencies to distribute funding, and predict successful collaborations. Locally, researchers within scientific fields can use quantitative studies to examine the current state of their community, and re-position themselves. Within the field of scientometrics, researchers can benefit from adapting a more theory-driven line of investigation, which combines quantitative findings with cognitive theories and explanation.

References

Anderson, A., McFarland, D., & Jurafsky, D. (2012). Towards a Computational History of the ACL: 1980-2008. *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, 13–21.

Axelrod, R. (1997). The Dissemination of Culture. *The Journal of Conflict Resolution*, *41*(2), 203–226.

Backstrom, L., Huttenlocher, D., Kleinberg, J., & Lan, X. (2006). Group formation in large social networks. *The 12th ACM SIGKDD International Conference*, 44–54.

Beller, S., Bender, A., & Medin, D. L. (2012). Should anthropology be part of cognitive science? *Topics in Cognitive Science*, *4*(3), 342–353.

Bentley, R. A. (2008). Random drift versus selection in academic vocabulary: An evolutionary analysis of published keywords. *PLoS ONE*, *3*(8).

Berger, J. A. & Heath, C. (2005). Idea Habitats: How the Prevalence of Environmental Cues Influences the Success of Ideas. *Cognitive Science*, *29*(2), 195–221.

Bergmann, T., Dale, R., Sattari, N., Heit, E., & Bhat, H. S. (2016). The Interdisciplinarity of Collaborations in *Cognitive Science*. *Cognitive Science*.

Bhat, H. S., Huang, L.-H., Rodriguez, S., Dale, R., & Heit, E. (2015). Citation Prediction Using Diverse Features. *3rd ICDM Workshop on Data Science and Big Data Analytics (DSBDA-2015) in IEEE ICDM '15*.

Bickerton, D. (2007). Language evolution: A brief guide for linguists. *Lingua*, *117*(3), 510–526.

Bird, A. (2007). What Is Scientific Progress? *Nous*, *41*(1), 64–89.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 7784.

Blei, D. M. & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*(1), 17–35.

Blei, D. M. & Lafferty, J. D. (2009). Topic Models. In A. N. Srivastava & M. Sahami (Eds.), *Text Mining: Classification, Clustering, and Applications* (pp. 71–94). Boca Raton, FL: CRC Press.

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

Bollen, J., Crandall, D., & Junk, D. (2014). From funding agencies to scientific agency. *EMBO reports*, 1–3.

Börner, K., Contractor, N., Falk-Krzesinski, H. J., Fiore, S. M., Hall, K. L., Keyton, J., Spring, B., Stokols, D., Trochim, W., & Uzzi, B. (2010). A Multi-Level Systems Perspective for the Science of Team Science. *Science Translational Medicine*, *2*(49), 1–5.

Bornmann, L. & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, *66*(11), 2215–2222.

Bourget, D. & Chalmers, D. J. (2014). What do philosophers believe? *Philosophical Studies*, *170*(3), 465–500.

Carroll, J. M. (2006). Human-Computer Interaction. In *Encyclopedia of Cognitive Science* (pp. 1–4). Chichester: John Wiley & Sons, Ltd.

Casella, G. & George, E. I. (1992). Explaining the Gibbs Sampler. *The American Statistician*, *46*(3), 167–174.

Chalmers, D. J. (2015). Why Isn't There More Progress in Philosophy? *Philosophy*, *90*(1), 3–31.

Christiansen, M. H. & Kirby, S. (2003). Language Evolution: The Hardest Problem in Science? In M. H. Christiansen & S. Kirby (Eds.), *Language evolution* (pp. 1–15). Oxford: Oxford University Press.

Cohen Priva, U. & Austerweil, J. L. (2015). Analyzing the history of *Cognition* using Topic Models. *Cognition*, *135*, 4–9.

Croft, W. & Cruse, D. A. (2004). *Cognitive Linguistics*. Cambridge: Cambridge University Press.

Cronin, B. (2004). Bowling alone together: Academic writing as distributed cognition. *Journal of the American Society for Information Science and Technology*, *55*(6), 557–560.

Cronin, B., Shaw, D., & La Barre, K. (2003). A cast of thousands: Coauthorship and subauthorship collaboration in the 20th century as manifested in the scholarly journal literature of psychology and philosophy. *Journal of the American Society for Information Science and Technology*, *54*(9), 855–871.

De Battisti, F., Ferrara, A., & Salini, S. (2015). A decade of research in statistics: A topic model approach. *Scientometrics*, *103*(2), 413–433.

Dediu, D. & de Boer, B. (2015). Language evolution needs its own journal. *Journal of Language Evolution*, lzv001.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Dietrich, E. (2011). There Is No Progress in Philosophy. *Essays in Philosophy*, *12*(2), 329–344.

Dong, Y., Johnson, R. A., & Chawla, N. V. (2015). Will this paper increase your *h*-index? In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15* (pp. 149–158). New York, New York, USA: ACM Press.

Evans, V., Bergen, B., & Zinken, J. (2007). The cognitive linguistics enterprise: An overview. In V. Evans, B. Bergen, & J. Zinken (Eds.), *The cognitive linguistics reader* (pp. 1–36). Equinox.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software*, *25*(5), 1–54.

Fitch, T. W. (2010). *The Evolution of Language*. Cambridge, MA: Cambridge University Press.

Foster, J. G., Rzhetsky, A., & Evans, J. a. (2015). Tradition and Innovation in Scientists' Research Strategies. *American Sociological Review*, *80*(5), 875–908.

Freeman, L. C. (1978). Centrality in Social Networks: Conceptual Clarification. *Social Networks*, *1*(3), 215–239.

Geeraerts, D. (2006). A rough guide to Cognitive Linguistics. In D. Geeraerts (Ed.), *Cognitive Linguistics: Basic Readings* (pp. 1–28). Berlin: Mouton de Gruyter.

Gelfand, A. E. (2000). Gibbs Sampling. *Journal of the American Statistical Association*, *95*(452), 1300–1304.

Gentner, D. (2010). Psychology in Cognitive Science: 1978-2038. *Topics in Cognitive Science*, *2*(3), 328–344.

Gersh, J. R., Mckneely, J. A., & Remington, R. W. (2005). Cognitive Engineering: Understanding Human Interaction with Complex Systems. *Johns Hopkins APL Technical Digest*, *26*(4), 377–382.

Gibbs, R. W. (1996). What's cognitive about cognitive linguistics? In E. H. Casad (Ed.), *Cognitive Linguistics in the Redwoods: The Expansion of a New Paradigm in Linguistics* (pp. 27–54). Berlin: de Gruyter.

Goldberg, A. E. (1995). *Constructions. A Construction Grammar Approach to Argument Structure*. Chicago: The University of Chicago Press.

Goldberg, A. E. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Goldstone, R. L. & Gureckis, T. M. (2009). Collective Behavior. *Topics in Cognitive Science*, *1*(3), 412–438.

Goldstone, R. L. & Janssen, M. A. (2005). Computational models of collective behavior. *Trends in Cognitive Sciences*, *9*(9), 424–430.

Goldstone, R. L. & Leydesdorff, L. (2006). The Import and Export of *Cognitive Science*. *Cognitive Science*, *30*(6), 983–993.

Goldstone, R. L., Roberts, M. E., & Gureckis, T. M. (2008). Emergent processes in group behavior. *Current Directions in Psychological Science*, *17*(1), 10–15.

Goldstone, R. L., Wisdom, T. N., Robert, M. E., & Frey, S. (2013). Learning Along With Others. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 58, pp. 1–45). Amsterdam: Academic Press.

Gonzales-Marquez, M., Mittelberg, I., Coulson, S., & Spivey, M. J. (Eds.). (2007). *Methods in Cognitive Linguistics*. Amsterdam: John Benjamins.

Granovskiy, B., Gold, J. M., Sumpter, D. J. T., & Goldstone, R. L. (2015). Integration of social information by human groups. *Topics in Cognitive Science*, *7*(3), 469–493.

Griffiths, T. L. & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 5228–5235.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211–244.

Grün, B. & Hornik, K. (2011). *topicmodels*: An R Package for Fitting Topic Models. *Journal of Statistical Software*, *40*(13), 1–30.

Guimera, R., Uzzi, B., Spiro, J., & Amaral, L. A. N. (2005). Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. *Science*, *308*(5722), 697–702.

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. *EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 363–371.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. New York: Springer.

Higgins, A. & Dyschkant, A. (2014). Interdisciplinary collaboration in philosophy. *Metaphilosophy*, *45*(3), 372–398.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572.

Huber, G. P. & Lewis, K. (2010). Cross-Understanding: Implications for Group Cognition and Performance. *Academy of Management Review*, *35*(1), 6–26.

Hyland, K. (1999). Academic attribution: Citation and the construction of disciplinary knowledge. *Applied Linguistics*, *20*(3), 341–367.

Jablonka, E. & Lamb, M. J. (2005). *Evolution in Four Dimensions.* Cambridge, MA: The MIT Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. New York: Springer.

Keen, I. (2014). Does cognitive science need anthropology? *Topics in Cognitive Science*, *6*(1), 150–151.

Kempe, D., Kleinberg, J., & Tardos, É. (2003). Maximizing the spread of influence through a social network. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data mining - KDD '03*, 137–146.

Kolaczyk, E. (2009). *Statistical analysis of network data*. New York: Springer.

Koschützki, D., Lehmann, K. A., Peeters, L., Richter, S., Tenfelde-Podehl, D., & Zlotowski, O. (2005). Centrality Indices. In U. Brandes & T. Erlebach (Eds.), *Network Analysis* (pp. 16–61). Berlin: Springer.

Kreuzman, H. (2001). A co-citation analysis of representative authors in philosophy: Examining the relationship between epistemologists and philosophers of science. *Scientometrics*, *51*(3), 525–539.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. The University of Chicago Press.

Lakatos, I. & Musgrave, A. (Eds.). (1970). *Criticism and the Growth of Knowledge*. Cambridge: Cambridge University Press.

Lakoff, G. (1990). The Invariance Hypothesis: Is abstract reason based on image-schemas? *Cognitive Linguistics*, *1*(1), 39–74.

Lakoff, G. (2004). *Don't Think of an Elephant!* White River Junction, VT: Chelsea Green Publishing.

Lakoff, G. & Johnson, M. (1980). *Metaphors we live by*. Chicago: The University of Chicago Press.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259–284.

Larivière, V., Haustein, S., & Börner, K. (2015). Long-distance interdisciplinarity leads to higher scientific impact. *PLoS ONE*, *10*(3), 1–15.

Laudan, L. (1977). *Progress and its Problems: Toward a Theory of Scientific Growth*. University of California Press.

Leydesdorff, L. (2001). *The Challenge of Scientometrics: The Development, Measurement, and Self-Organization of Scientific Communications*. Leiden, the Netherlands: DSWO Press.

Leydesdorff, L. & Goldstone, R. L. (2014). Interdisciplinarity at the journal and specialty level: The changing knowledge bases of the journal *Cognitive*

Science. *Journal of the Association for Information Science and Technology*, *65*(1), 164–177.

Leydesdorff, L. & Milojević, S. (2015). Scientometrics. *International Encyclopedia of the Social & Behavioral Sciences*, *21*, 322–327.

Losee, J. (2004). *Theories of Scientific Progress: An Introduction*. Routledge.

Lucas, C., Nielsen, R. a., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis*, 1–24.

Mao, A., Mason, W., Suri, S., & Watts, D. J. (2016). An Experimental Study of Team Size and Performance on a Complex Task. *PLOS ONE*, *11*(4), e0153048.

Marx, W., Bornmann, L., Barth, A., & Leydesdorff, L. (2014). Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS). *Journal of the Association for Information Science and Technology*, *65*(4), 751–764.

Mason, W. & Watts, D. J. (2012). Collaborative learning in networks. *Proceedings of the National Academy of Sciences*, *109*(3), 764–769.

Milojević, S. (2014). Principles of scientific research team formation and evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(11), 3984–9.

Milojević, S. (2015). Quantifying the cognitive extent of science. *Journal of Informetrics*, *9*(4), 962–973.

Mimno, D. (2012). Computational Historiography: Data Mining in a Century of Classics Journals. *Journal on Computing and Cultural Heritage*, *5*(1), 3:1–3:19.

Moore, D. S. (2015). *The Developing Genome: An Introduction to Behavioral Epigenetics*. Oxford University Press.

Moussaid, M., Garnier, S., Theraulaz, G., & Helbing, D. (2009). Collective information processing and pattern formation in swarms, flocks, and crowds. *Topics in Cognitive Science*, *1*(3), 469–497.

Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(2), 404–409.

Nichols, L. G. (2014). A topic model approach to measuring interdisciplinarity at the National Science Foundation. *Scientometrics*, *100*(3), 741–754.

Nielsen, K. (1987). Can there be progress in Philosophy? *Metaphilosophy*, *18*(1), 1–30.

Niiniluoto, I. (1980). Scientific progress. *Synthese*, *45*, 427–462.

Oldham, P., Hall, S., & Burton, G. (2012). Synthetic Biology: Mapping the Scientific Landscape. *PLoS ONE*, *7*(4), e34368.

Petersen, A. M. (2015). Quantifying the impact of weak, strong, and super ties in scientific careers. *Proceedings of the National Academy of Sciences*, *2015*, E4671–E4680.

Pons, P. & Latapy, M. (2005). Computing Communities in Large Networks Using Random Walks. In P. Yolum, T. Güngör, F. Gürgen, & C. Özturan (Eds.), *Computer and Information Sciences - ISCIS 2005* (Vol. 3733, pp. 284–293). Lecture Notes in Computer Science. Berlin: Springer.

Ponweiser, M. (2012). *Latent Dirichlet Allocation in R* (Phd Thesis, Vienna University of Economics and Business).

Porter, A. L., Cohen, A. S., David Roessner, J., & Perreault, M. (2007). Measuring researcher interdisciplinarity. *Scientometrics*, *72*(1), 117–147.

Porter, M. F. (2001). Snowball: a language for stemming algorithms. `http://snowball.tartarus.org/texts/introduction.html`.

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S., & Carey, T. (1994). *Human-computer interaction*. Essex, UK: Addison-Wesley Longman Ltd.

R Core Team. (2016). *R: A Language and Environment for Statistical Computing*. `https://www.R-project.org/`. R Foundation for Statistical Computing. Vienna, Austria.

Regier, T. (1996). *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. Cambridge, MA: The MIT Press.

Rinia, E. J., van Leeuwen, T. N., Bruins, E. E. W., van Vuren, H. G., & van Raan, A. F. J. (2002). Measuring knowledge transfer between fields of science. *Scientometrics*, *54*(3), 347–362.

Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013). The structural topic model and applied social science. *NIPS 2013 Workshop on Topic Models*, 2–5.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. *Proceedings of the 20th Conference on Uncertainty in Artificial intelligence*, 487–494.

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science*, *311*(5762), 854–856.

Schelling, T. C. (1971). Dynamic models of segregation. *The Journal of Mathematical Sociology*, *1*(2), 143–186.

Shi, X., Leskovec, J., & McFarland, D. a. (2010). Citing for high impact. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries - JCDL '10* (pp. 49–58). ACM Press.

Shin, J. C., Toutkoushian, R. K., & Teichler, U. (Eds.). (2011). *University rankings: Theoretical basis, methodology and impacts on global higher education*. Dordrecht: Springer.

Sinha, C. (2007). Cognitive linguistics, psychology and cognitive science. In D. Geeraerts & H. Cuyckens (Eds.), *Handbook of Cognitive Linguistics* (pp. 1266–1294). Oxford: Oxford University Press.

Steyvers, M. & Griffiths, T. (2007). Probabilistic Topic Models. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 427–448). Hillsdale, NJ: Lawrence Erlbaum.

Swan, A. (2007). Open access and the progress of science. *American Scientist*, *95*, 197–199.

Theiner, G., Allen, C., & Goldstone, R. L. (2010). Recognizing group cognition. *Cognitive Systems Research*, *11*(4), 378–395.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical Combinations and Scientific Impact. *Science*, *342*(6157), 468–472.

Wang, X. & McCallum, A. (2006). Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '06* (pp. 424–433). New York, New York, USA: ACM Press.

Watson, J. D. & Crick, F. H. C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, *171*(4356), 737–738.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330*(6004), 686–688.

Wray, K. B. (2006). Kuhnian Revolutions Revisited. *Synthese*, *158*(1), 61–73.

Wray, K. B. & Bornmann, L. (2015). Philosophy of science viewed through the lense of "Referenced Publication Years Spectroscopy" (RPYS). *Scientometrics*, *102*(3), 1987–1996.

Yan, R., Tang, J., Liu, X., Shan, D., & Li, X. (2011). Citation count prediction: Learning to estimate future citations for literature. *CIKM '11 Proceedings of the 20th ACM international conference on Information and knowledge management*, 1247–1252.

Yang, T.-I., Torget, A. J., & Mihalcea, R. (2011). Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 96–104).

Yau, C. K., Porter, A. L., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, *100*(3), 767–786.