The Intellectual Organisation of History

THÈSE N^{O} 8537 (2018)

PRÉSENTÉE LE 28 AOÛT 2018

AU COLLÈGE DES HUMANITÉS
LABORATOIRE D'HUMANITÉS DIGITALES
PROGRAMME DOCTORAL EN MANAGEMENT DE LA TECHNOLOGIE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Giovanni COLAVIZZA

acceptée sur proposition du jury:

Prof. D. Foray, président du jury
Prof. F. Kaplan, Prof. M. Franceschet, directeurs de thèse
Prof. C. Sugimoto, rapporteuse
Prof. R. Bod, rapporteur
Prof. G. de Rassenfosse, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2018

The intellectual organization of history

Giovanni Colavizza

“Man is unique not because he does science, and he is unique not because he does art, but because science and art equally are expressions of his marvellous plasticity of mind.” Jacob Bronowski, The Ascent of Man

Acknowledgements

Someone wise and kind once told me that writing the acknowledgements has been her thesis best moment. I cannot but agree now: it is a source of genuine pleasure to stop and think of all the persons that helped me, taught me something, gave their support over these years. The list is longer, the debt larger than I am capable of publicly state here, yet I have done my best to recognize the most important influences on this endeavor of mine.

I would like to start by thanking my thesis directors, Frédéric and Massimo: we make for the most awkward of teams, and yet they both helped and advised a great deal in their own ways, but most importantly gave me the freedom to make this adventure my own. The Venice Time Machine and Linked Books projects, which provided the framework for this thesis, have been an institutional collaboration among the École Polytechnique Fédérale de Lausanne (EPFL), the Ca'Foscari University of Venice and its Humanities Library, the State Archive of Venice, the *Marciana* Library, the *Istituto Veneto di Scienze, Lettere ed Arti*, the European Library of Information and Culture in Milan, the Union Catalogue of Italian Libraries and Bibliographic Information in Rome. I thank all for their support.

Several persons helped me in this journey. I would like to thank in particular the VTM team in Venice (in no particular order): Martina Babetto, Silvia Ferronato, Andrea Erboso, Fabio Bortoluzzi, Francesca Zugno and Davide Drago for their work and the many shared moments over the years. Thanks also to Daniela Grandin, for her incredible support and enthusiasm, to Dorit Raines, for being such an unlimited source of energy, wit and good advice, and to Mario Infelise, for the most enriching of conversations and for being always inquisitive and uncompromising, wisely and where appropriate. I have the most fond memories of many colleagues at EPFL with whom I shared the past years, in particular our beloved Maud Ehrmann, the true Orlin Topalov and Alicia Foucart Noriega, who is always kind and willing to help. I am grateful to Cristina Dondi and her magnificent group in Oxford for their kindness. My time spent at the Centre for Science and Technology Studies in Leiden was terrific, a way too short experience where I met a fantastic group of people. Thanks to the great Ludo Waltman, who is never tired of listening to my ever-changing ideas, offering rigor and enthusiasm in balance, and to Vincent Traag, Nees Jan van Eck, Thomas Franssen, Thed

van Leeuwen and really all of CWTS. Many more persons have enriched my journey in a variety of ways, whom I thank without naming.

I am most indebted to my friend Matteo Romanello, not least for his endless patience with me: we worked together on this project and I could not have made it without him. To Laurent Bolli, also friend in life and business, for so many things, but especially for opening my eyes to design and beauty: I still have a long way to go. With both, the journey has just begun.

I thank my family for its support and unwavering love, in particular my mother Anna Pia, and all my friends old an new, for sharing this life together.

My work is dedicated to Stefania and to the memory of my father Baldo.
What you gave me I cannot express in words. I love you.

Abstract

A tradition of scholarship discusses the characteristics of different areas of knowledge, in particular after modern academia compartmentalized them into disciplines. The academic approach is often put to question: are there two or more cultures? Is an ever-increasing specialization the only way to cope with information abundance or are holistic approaches helpful too? What is happening with the digital turn? If these questions are well studied for the sciences, our understanding of how the humanities might differ in their own respect is far less advanced. In particular, modern academia might foster specific patterns of specialization in the humanities. Eventually, the recent rise in the application of digital methods to research, known as the digital humanities, might be introducing structural adaptations through the development of shared research technologies and the advent of organizational practices such as the laboratory. It therefore seems timely and urgent to map the intellectual organization of the humanities. This investigation depends on few traits such as the level of codification, the degree of agreement among scholars, the level of coordination of their efforts. These characteristics can be studied by measuring their influence on the outcomes of scientific communication. In particular, this thesis focuses on history as a discipline using bibliometric methods.

In order to explore history in its complexity, an approach to create collaborative citation indexes in the humanities is proposed, resulting in a new dataset comprising monographs, journal articles and citations to primary sources. Historians' publications were found to organize thematically and chronologically, sharing a limited set of core sources across small communities. Core sources act in two ways with respect to the intellectual organization: locally, by adding connectivity within communities, or globally as weak ties across communities. Over recent decades, fragmentation is on the rise in the intellectual networks of historians, and a comparison across a variety of specialisms from the human, natural and mathematical sciences revealed the fragility of such networks across the axes of citation and textual similarities. Humanists organize into more, smaller and scattered topical communities than scientists.

A characterization of history is eventually proposed. Historians produce new historiographical knowledge with a focus on *evidence* or *interpretation*. The former aims at providing the community with an agreed-upon factual resource. Interpretive work is instead mainly focused on creating novel perspectives. A second axis refers to two modes of exploration of new ideas: *in-breadth*, where novelty relates to adding new, previously unknown pieces to the mosaic, or *in-depth*, if novelty then happens by improving on previous results. All combinations possible, historians tend to focus on in-breadth interpretations, with the immediate consequence that growth accentuates intellectual fragmentation in the absence of further consolidating factors such as theory or technologies. Research on evidence might have a different impact by potentially scaling-up in the digital space, and in so doing influence the modes of interpretation in turn. This process is not dissimilar to the gradual rise in importance of research technologies and collaborative competition in the mathematical and natural sciences. This is perhaps the promise of the digital humanities.

Keywords: Bibliometrics, Scientometrics, Citation indexes, Science mapping, Network science, Digital humanities, Humanities, History, History of Venice, Venice.

Sommario

Un'intera tradizione di studi considera le caratteristiche di differenti approcci conoscitivi, in particulare da quando l'accademia moderna li ha distinti in altrettante discipline. Questa organizzazione è spesso messa in discussione: esistono realmente due o più distinte culture della conoscenza? Possiamo gestire la massa di letteratura scientifica solamente grazie alla specializzazione disciplinare o esistono delle alternative con meriti loro propri? Cosa comporta il digitale in tutto ciò? Se queste domande sono abitualmente considerate riguardo alle scienze naturali e matematiche, lo sono molto meno rispetto alle scienze umane. La strutturazione dell'accademia moderna probabilmente ha effetti diversi sulle scienze umane dal punto di vista della loro organizzazione intellettuale. L'avvento dell'informatica umanistica, o *digital humanities*, a sua volta sta cambiando le cose tramite lo sviluppo di tecnologie di ricerca condivise e l'introduzione di pratiche organizzative proprie di altre discipline, come il laboratorio. Il momento è dunque propizio per studiare come le scienze umane sono organizzate intellettualmente, ora e nel passato prossimo, considerando aspetti quali la codificazione della conoscenza prodotta o il livello di accordo e coordinazione tra studiosi. Questi elementi possono essere studiati misurandone gli effetti sui prodotti della comunicazione scientifica, quali le pubblicazioni. Questo lavoro di tesi considera il caso di studio della storiografia e l'uso di metodi bibliometrici a questo scopo.

Visto l'obiettivo di studiare la storiografia nella sua complessità, un primo contributo riguarda un nuovo approccio per creare indici di citazioni nelle scienze umane, in maniera distribuita e collaborativa. A seguito dell'applicazione di questo approccio ad un caso di studio specifico, un nuovo insieme di dati è stato creato, per permettere di considerare citazioni su piani distinti ed interrelati quali monografie, articoli di rivista e fonti primarie. Uno studio esplorativo ha poi permesso di evidenziare come gli storici si organizzino perlopiù per aree tematiche o cronologiche, in molte e piccole comunità che condividono poche fonti in comune, sia primarie che secondarie. In particolare, nel ridotto numero di fonti principali, vale a dire molto citate, troviamo solamente monografie e opere di riferimento quali edizioni critiche, riconosciute e citate globalmente, al di là di ristrette comunità. In generale, la letteratura sia primaria che secondaria, rimane utilizzata solo localmente, da pochi studiosi. Negli ultimi decenni, questa già precaria organizzazione intellettuale risulta ulteriormente in frammentazione. Se confrontata con altri ambiti disciplinari nelle scienze naturali e matematiche, l'organizzazione intellettuale degli umanisti risulta composta da molte, piccole comunità con poco in comune tra loro.

Questo lavoro propone infine una nuova caratterizzazione della storia come disciplina, considerata attraverso le lenti della storiografia. Gli storici producono lavori che possono essere caratterizzati da una attenzione privilegiata per l'evidenza storica o per l'interpretazione di fatti storici. I lavori sull'evidenza, quali le edizioni critiche, si orientano a creare delle risorse condivise per la comunità, mentre quelli interpretativi hanno come obiettivo principale la discussione di nuove ed originali prospettive su tematiche storiche. Gli storici adottano inoltre due approcci nel loro lavoro, procedendo in profondità su una fonte o un tema, o estensivamente quando cercano di coprire un territorio di ricerca nella sua ampiezza. Se nel primo caso il lavoro dello storico approssima quello cumulativo di molte scienze naturali, nel secondo esso si concentra sulla novità di una tematica o una prospettiva non prima considerata. L'attuale comunità di storici si orienta perlopiù su lavori interpretativi che procedono in estensione, non a caso i più prestigiosi. Tuttavia, in questa tesi emerge chiaramente l'impatto dei contributi in profondità sull'evidenza primaria, che influenzano a lungo generazioni di storici. Con l'avvento dell'informatica umanistica e del digitale anche nell'ambito storico e umanistico, è possibile che nuove tecnologie di ricerca e modi di collaborare cambino profondamente la tradizionale organizzazione intellettuale e sociale delle scienze umane.

Parole chiave: Bibliometria, Scientometria, Indici di citazioni, Mappe delle scienza, Scienza delle reti, Informatica umanistica, Scienze umane, Storia, Storia di Venezia, Venezia.

Contents

1 Introduction

The humanities are nowadays considered as a galaxy of disciplines part of the organization of research and higher education. The institutional separation of the humanities and the sciences was gradual and debated, becoming pronounced only in modern academia during the 19^{th} century [Bouterse and Karstens, 2015].[1] Indeed common roots, mainly attributable to philology [Turner, 2015] and shared methodological patterns [Bod, 2013, 2018], are strong even within the humanities. If the humanities have not been exempt from the push to divide, specialize and conquer which is one of the main traits of modern academia [Wellmon, 2015], they have perhaps done it in their distinct way. More generally, the condition of the humanities within academia seems to betray the uneasiness of something which is forced into a condition not its own. While the sciences thrived during the past centuries, their moments of crisis being morally or socially grounded first and foremost (e.g. Luddism or the debate around the atomic bomb), the humanities have endured a seemingly everlasting state of crisis and self-doubt [Plumb, 1964]. Still well-known is the Rede lecture given by C. P. Snow in Cambridge [Snow, 1959], where he lamented the apparently insurmountable gulf existing between scientists and 'literary intellectuals' – the two cultures – made of mutual incomprehension at times bordering into disdain. The tendency of contemporary academia to organize into a "one size to fit them all" is still a source of constant worry when considering topics as important as education, funding and research evaluation [Reale et al., 2017].

The organization and relations among different areas of knowledge, as well as their demarcation, are long-lasting topics of discussion. An early attempt to establish a hierarchy of the sciences dates to August Comte, who proposed a ranking starting from mathematics and ending with sociology, organized according to the raising complexity of the subject matter (mathematics being the least complex) and declining generality of results (mathematics being the most general). Sciences also build on top of each other, with sociology as the predictive science of society necessitating contributions from the rest of

[1] In what follows we distinguish between the human and natural or mathematical sciences simply from an institutional point of view, merely operationally. "Humanities" is often used as a shorthand for human sciences, as "sciences" is used as a shorthand for natural or mathematical sciences. Further defining these terms is beyond the point, as one of the goals of this work is indeed to explore if, what is within these general categories, differs under some measurable aspects.

the sciences [Bourdeau, 2015]. During the 19^{th} century still, the dichotomy between the sciences and the humanities was also grounded in philosophical discourse. Most notably, Wilhelm Dilthey advanced a distinction between the natural and human sciences grounded in two distinct approaches to knowledge: a focus on explanation and cause-effect relations for the former, on understanding and part-whole relations for the latter [Dilthey, 1883]. Further to that, Wilhelm Windelband distinguished between *nomothetic* and *idiographic*, or a tendency to generalize explanations into laws and a tendency to specify interpretive insights on the particular, respectively [Windelband, 1904]. Post-positivist thinking during the 20^{th} century has offered, from a multiplicity of perspectives, a different, sometimes opposite view on the possibility of a hierarchy of the sciences, often suggesting no structure is to be found and emphasizing the irrational, contingent or constructive side of this phenomenon (e.g. Dupré [1983]; Feyerabend [2010]). Within these two extremes, recent work has found bibliometric evidence for a hierarchy of the sciences suggesting that different intellectual organizations might indeed exist [Fanelli and Glänzel, 2013], while a tradition of sociological studies has shown how the disciplines part of the sciences and the humanities can differ substantially in terms of their social and institutional organization (e.g. Whitley [1984]). From the point of view of the history of science and the humanities, increasing amount of work is reconsidering and contextualizing the development of disciplinary boundaries [Krämer, 2018]. These findings encourage the gradual extension of studies aiming at individuating and qualifying commonalities and differences across all the sciences. As a different outcome from the same process is sometimes a sign of a difference in the inputs of the experiment, so the current state of the disciplines part of the sciences and the humanities, when compared within a sociological and bibliometric framework, might cast new light on the separation between the two. It is then within an historical setting that it would be possible to further clarify how the current state of affairs came to be.

Bibliometrics is the area of research concerned with the statistical analysis of publications, typically but not exclusively scholarly ones, and is strongly related to science studies in general and the study of scientific communication in particular [De Bellis, 2009]. For historical and practical reasons, bibliometrics developed a focus on the sciences: their (citation) indexation, study and evaluation were considered in successive moments each building on the results of previous ones. The humanities were and still are mostly considered through the same lenses that were developed for the sciences, even if sometimes little or even contrary evidence in support for such a choice is actually available [Ardanuy, 2013]. This is particularly problematic with respect to research evaluation, where persisting failure to properly take into account SSH (social sciences and humanities) features is recognized [Reale et al., 2017], so much so that alternative approaches are by now pursued [Hammarfelt, 2016]. Efforts into basic research exploring the intellectual organization of the humanities cannot but help to overcome these limitations, especially as this is a time when research and scholarly communication practices in the humanities are undergoing significant changes due to the digital turn.

There is a relatively long tradition of applying digital methods to study the objects investigated in the humanities, which broadly goes under the name of "humanities computing" [McCarty, 2003]. The term "digital humanities" is instead more recent, having been introduced by Schreibman et al. [2004]. It stands as an effort to broaden the scope of the approach to include the use of traditional humanities methods to study digital objects, and explicitly stress the need to go beyond 'mere digitization'. Furthermore, the digital humanists have taken the lead to expand the traditional means of scholarly communication used by humanists, to experiment more broadly with the Web and to include, among others, new social media, blogs, crowdsourcing, open access and alternative forms of peer review. Despite the fact that the digital humanities are now broader, richer and more varied, and their efforts are fostering interdisciplinarity and collaborations [McCarty, 2015], there still seems to be a long way to go before their full potential can fully manifest. Digitized and born digital data or artefacts, gradually accumulating and becoming interlinked, might determine a profound shift in research approaches and thus in the intellectual and social organization of the humanities. The digital humanities are in particular starting to bring organizational practices such as collaboration, teamwork, laboratories and larger projects with more funding which, if they are commonplace in some sciences, are quite alien to the humanities. These phenomena still unfolding, their consequences all too open ended, it seems all the more urgent to better comprehend the intellectual organization of humanists and what might be changing due to the digital turn or other concurring trends.

1.1 Aims of the thesis

This thesis offers an investigation into the intellectual organization of history, a discipline (contentiously) part of the humanities, using methods from bibliometrics and science and technology studies (STS), and theoretical frameworks borrowed from the philosophy and sociology of science. The intellectual organization, or cognitive structure of a certain unit of analysis in science – a discipline or, more likely, research areas of different size and scope within it – covers the way knowledge is created, related and new one accumulates from the point of view of its contents. The intellectual organization is therefore shaped by the interactions of its structural configuration, or the landscape, with the dynamic accumulation and integration of new knowledge with existing one. It is also evidently connected to, yet distinct from the social organization of the same unit of analysis, since the objects of study usually differ – publications and people, respectively. The intellectual organization is instead very close to, and at times blurred with scientific communication, as it is through these products that the intellectual organization is typically measured in some of its aspects. As it will be discussed later on, the intellectual organization of scientific fields has been amply considered as an object of study in previous literature. In fact, we might say that the approach taken here is more akin to STS, as the starting point and focus of analysis are the conceptual aspects of science, and not so much its social organizational ones [Leydesdorff, 1989]. In this work, the intellectual organization will be studied bibliometrically via citations and the full text of scholarly publications, considered as quantifiable objects which can be used to put publications into relation within a network representation. This approach is typically taken in science mapping. To be sure, not all the aspects of the intellectual organization of a unit of analysis in science are quantifiable, and there are indeed ways to capture some of the non-quantifiable ones using qualitative methods, which are nevertheless beyond the scope of the present work.

The principal aim of the thesis is thus to develop and apply bibliometric and STS methods to study the intellectual organization of the humanities, with a focus on history. We are in part motivated by the still substantial lack of such efforts in the domain, which in part hinders the appropriate consideration of the humanities in bibliometrics, including from an evaluative point of view. Furthermore, it is felt that the humanities might be at the fringe of significant changes in their intellectual and social practices, due to the more general increasing digitization and "datafication" of science: it seems thus timely to map their intellectual landscapes. Precondition to this is finding novel ways to expand the bibliometric data coverage for the humanities, which is currently much lacking. Furthermore, another goal is proposing an attempt to connect, or reconnect, bibliometrics with the history and sociology of science by setting the development and application of bibliometric methods into a theoretical framework whenever possible. In the former case, the use of bibliometrics for historical research has been proposed and at times attempted but has been so far very little explored, with close to no impact on the history of science [Hérubel, 1999; Garfield et al., 2003]. In the latter case instead, the studies in the sociology of science taking a bibliometrics turn have a richer tradition, now-a-days less central within the field and especially so with respect to the humanities [Franssen and Wouters, 2017], despite the persisting importance of theoretical developments in bibliometrics [Sugimoto and Cronin, 2016]. More specifically, there seems to be a divide between sociological studies taking a qualitative approach and bibliometrics studies taking a quantitative one, which has set apart the two communities [Hammarfelt, 2012a]. A fourth and last goal is attempting to understand the potential impact of recent developments, mainly the current digital turn, which is influencing many dimensions of society including scholarship in the humanities.

In summary, this thesis considers the following overarching questions:

1. Can we enlarge the bibliometric data coverage for the humanities? A precondition for any bibliometric investigation is the availability of the necessary dataset to empirically answer a certain question. Indeed, one of the main hindrances to the advancement of bibliometrics for the SSH has been the lack of coverage in citation indexes and digitally available publication corpora.

2. How can we bibliometrically represent and study research fields in the humanities, with a focus on their intellectual organization and informed by a theoretical framework? This second question develops two related aspects: one methodological, the other theoretical. The methodological aspect considers the development of dedicated methods to represent and analyze the intellectual organization of fields in the humanities, the second aspect casts a theoretical view on the development and application of such methods, especially when advancing hypotheses and interpreting results.

3. Can we individuate structural elements in such intellectual organization? The third question considers the applications of methods on specific case studies, for the purpose of individuating structural elements which can be observed in order to detect ongoing trends into the

intellectual organization of research fields. An example are core highly cited sources, and the role they play into connecting different areas of an intellectual landscape.

4. How is new knowledge accumulated within such intellectual organization(s) and how is this changing? The last question blends previous ones together, considering a process, that of knowledge accumulation or the integration of new scholarly results into the existing body of knowledge. This process can be considered over short or long amounts of time, in order to detect and discuss trends.

To this date, research efforts along these lines have been substantially hindered by two mutually reinforcing factors: the intrinsic complexity of communication practices in the humanities, which include but are not limited to citation behavior, types of outputs and heterogeneity of audiences, and the lack of bibliometric data [Hellqvist, 2009]. It is well known that humanists use a spectrum of publication typologies which is varied and likely complementary, such as monographs, edited volumes, journal articles. Their interplay is still relatively poorly studied. Furthermore, they make reference to a broad variety of sources, besides secondary literature. Most notably, humanists refer to primary sources in footnotes, a category which, in the case of history, usually includes archival documents and can span pretty much anything that might be relevant. Eventually, referencing practices and citation behaviors can be quite complex, both syntactically and semantically [Grafton, 1999; Zerby, 2003]. All these challenges, and the fact that historically citation indexes focused on journal articles due to their importance for the (other) sciences, makes the coverage of the humanities within bibliometric databases quite poor [Mongeon and Paul-Hus, 2016]. Not only journal articles are a secondary albeit important publication typology for the humanities, but no citation index to this date ventures into indexing primary sources. Despite the fact that little can be done with respect to the former set of challenges, if not to acknowledge them, some consideration will be given to the latter, where the state of the art can be advanced.

1.2 Bibliometrics

The research questions and goals of this thesis could be approached from a variety of perspectives, including history and sociology of science. The choice of bibliometrics needs to be somewhat justified (for what follows see

De Bellis [2009]). The origins of the application of quantitative or statistical methods to analyze publications are related to library catalogs and acquisition policies during the first decades of the past century. The accumulation of scholarly literature fostered, during the 1920s and 30s, the development of some fundamental laws which apply also to bibliometrics: Lotka's, Bradford's and Zipf's. It also led to the realization that the amount of published scholarship was rapidly going out of control, thus mechanized solutions were called for [Bush, 1945]. It is from this need that one pivotal innovation was introduced by Eugene Garfield in 1960: the *Science Citation Index*. The first and most important citation index offered not only an information retrieval tool, but also an object for self-inspection: a way to turn the tools of science onto itself [De Solla Price, 1986].

The field which can nowadays perhaps be called science of science, originates from an interest into the quantitative study of science as a social and information phenomenon (scientometrics) and the widespread use of bibliometric data for its advancement, such as publications and citations. Yet the convergence of interests during the 1960s and 70s was broader than that and is perhaps better viewed from today's eye as the origins of science studies. Pivotal contributions were made from the philosophy of science (among others Kuhn's Structure of Scientific Revolutions first edition in 1962 [Kuhn, 1996]) and the sociology of science (e.g. Merton's study of the Matthew Effect and his collected essays on the Sociology of Science [Merton, 1968, 1974]), to name but two fields. Nowadays the sociology of science is largely split in two groups, roughly a quantitative community and a qualitative one, a divide also felt in the bibliometrics community between who argues for a more theoretically informed approach and who claims that theory is of little practical help [Hammarfelt, 2012a].

The field of bibliometrics underwent a turn of interests since the 1980s towards the development of methods for the evaluation of research and their application on bibliometric data [Moed et al., 1985]. Since then, the field has strongly focused on impact evaluation, with less attention given to the seminal links with the sociology of science [Franssen and Wouters, 2017]. The relation with library and information studies has instead remained stable, mainly via empirical works for the purpose of collection development or other internal needs.

The principal motivation for taking a bibliometrics and STS approach in the present study is that it is felt that a quantitative approach can yield insights and knowledge that are difficult to obtain qualitatively, with tools such as surveys and interviews. More specifically, data can be sampled at a certain representative scale in order to capture the overall features of a system, albeit in a simplified form, instead of attempting to reconstruct it from within. It is also felt that the best empirical test for any theory requires using quantitative methods and representative data at scale, despite the fact that this operation usually entails the need to operationalize the theory, if this latter is not expressed formally to begin with (as is rarely the case in the sociology of science). This work is thus an explicit effort to reconnect bibliometrics with science studies. Yet the present study could and should be complemented by efforts taking alternative routes.

1.3 History

Institutionally, the humanities are nowadays a collection of academic disciplines, whose list can change according to the system under consideration. As anticipated above, it is not before the 19^{th} century that such separation began to be actively discussed, including advancing some motivations for its existence. There exist a variety of options on how to define and operationalize the category of "discipline", or any smaller unit of analysis within. Sugimoto and Weingart [2015] suggest that the discipline category can be conceptualized from the following perspectives: cognitive (focusing on contents, theories and methods), social, communicative (discourses and jargon), separatedness of the body of knowledge, tradition and institutional (e.g. affiliation). The perspective taken here focuses on the related cognitive and communicative aspects of a certain unit of analysis, as determinants and signals of its intellectual organization. Furthermore, the same authors suggest that operationalizations of the aforementioned perspectives can focus on publications, people and ideas. We strongly focus on publications both to bound a unit of analysis (e.g. via the venues of publication) and to measure specific aspects of it (e.g. via citations), and on ideas via work on the full text of such publications in an attempt to map the language and jargon in use. The people category will only be considered in passing, for example regarding co-authorships, since the focus is primarily on scientific contents and the body of knowledge produced by historians and not directly on them.

With respect to the units of analysis under consideration, we will only episodically move outside of a single discipline, and favour instead smaller and more homogeneous units of analysis. This choice is motivated by the rising awareness that disciplinary boundaries, paralleled by the need to overcome them explicitly via interdisciplinary efforts, might be useful for the administration of science but less so for its development and study. Furthermore, data-driven or bottom-up approaches have proven to be more accurate than top-down or platonic conceptualization of disciplinary boundaries in practical settings such as the classification of scientific outputs [Klavans and Boyack, 2016]. We therefore make the following terminological choices: we use "discipline" for broad categories with a substantial institutional flavor (such as history or literature); we use instead "specialism" to indicate a more focused unit within a discipline, which can be operationalized in a variety of ways, typically considering topics or objects of interest (e.g. history of Venice or economic history, but also economic history of Venice). From a people/publications point of view, a specialism is made of a broad but connected community of scholars, with few clear prominent ones, recognized shared venues of publication and conferences [Morris and Van der Veer Martens, 2008]. Considering ideas instead, a specialism shares methods, topics and questions. Eventually, we use the terms "field" or "research area" interchangeably for anything in-between a discipline and a specialism, when a third category is needed.

Given the need to bound the scope of the thesis, our focus will be on history. This might be considered a slight oddity, since history is sometimes viewed as bordering both the humanities and social sciences [Katz, 1995], yet precisely this aspect makes the choice particularly compelling. There exist some recent doctoral dissertations which take a bibliometric view on other disciplines part of the humanities, including literary studies [Hammarfelt, 2012a] and classics [Romanello, 2015], whilst history has not yet been considered from this perspective, thus this thesis begins to fill the gap. The scope of history as a discipline is very broad, encompassing the human past (in itself a potentially all-encompassing category) and the use of any evidence which is deemed relevant by the historian in order to convey a reasonable and plausible reconstruction of the phenomenon under consideration [Carr, 1998]. The in-principle not bounded scope of primary sources is one feature of history which makes it interesting to study. Over time, historians have worked more or less actively on their sources in order to make them available to other scholars, notably during the second half of the 19^{th} century (positivism). Primary sources are also made available to historians by the work of other professionals, for example archivists. The open ended scope of the historian's sources coupled with the fact that some of them are more accessible than others, makes history a compelling case to study the role and impact of the accessibility of primary evidence into the intellectual organization of the discipline. Lastly, as historians cover much ground, they also apply a broad variety of methods, according to their evidence and period of study, school, questions, preference. This variety allows to explore which of these factors, or others, plays a prominent role into the intellectual organization of the community.

1.4 Structure of the thesis

The thesis follows a progression from the specific to the generic in terms of datasets used, with the aim of broadening the generality of results. We will thus start by considering a specialism in history at a fine grained level, the history of Venice, to then use increasingly broader datasets from the main bibliometric databases. The thesis also follows a logical development into the subsequent steps of: data acquisition, exploration and mapping, measurement of specific phenomena of interest, broadening of scope by considering diachronic trends and comparisons with different disciplines. These steps correspond to separate chapters, and again attempt to move from the specific to the general in studying the intellectual organization of historians by posing increasingly broader questions.

Chapter 2 offers an overview of the relevant state of the art, with a focus on the bibliometrics literature considering the humanities, and history in particular. It also discusses some sociological frameworks which will be used to either guide the empirical research or discuss results in what follows, more specifically introducing Whitley's, Fuchs' and Becker and Trowler's works. The state of the art is rather comprehensive, also in order to lighten each individual chapter where only contextual or technical references are given.

The following Chapter 3 presents an approach for the citation indexation of scholarly literature in the humanities. Originally developed for the purpose of digitizing and indexing all the scholarly literature on the history of Venice relying on the collections of research libraries, it became a more general set of guidelines and a software platform to potentially replicate and extend the same work covering other specialisms or entire fields. The proposed approach has the merit of considering all cited sources, including primary ones, and of allowing to scale via collaborative efforts. The chapter discusses how to select a representative corpus of publications using library resources, presents a technical pipeline to build a citation index from digitized publications, and finally unveils two web interfaces, a digital library and a citation index, which can be used to access the data in a friendly way. Nevertheless, the main purpose of this work in the present context was to provide for a novel dataset to be used to explore the intellectual organization of a specialism in history at an unprecedented level of detail.

Chapter 4 starts by offering an exploratory science mapping study using some citation data from the Venetian dataset, in particular a dataset of monograph to monograph citations. The goal is twofold: to understand how this specialism organizes into clusters at a citation level rarely considered in the literature, that of monographs, and to explore the so called core sources: highly cited works which play a disproportionate role into connecting the bibliographic coupling representation of the specialism and shaping its landscape. This chapter further gives empirical motivation to pose a set of questions which will be thoroughly explored in the following chapters.

In the following Chapter 5, the quest to understand the structural role of core sources with respect to the intellectual organization of a specialism or entire field is explicitly brought forward. Core sources are likely important in the process of knowledge accumulation, as they might allow for ideas, perspectives, results and methods to emerge beyond the possibly narrow scope of specialized topics of discussion to instead reach a wide part of the community. Core sources also play an important if sometimes under-acknowledged role in several sociological theories dealing with the intellectual organization of scientific fields. This chapter thus introduces a methodological framework to study the structural role of core sources, and applies it to three datasets of increasing size and generality and different compositions in terms of the publication typologies they include.

The last two chapters move to consider broader topics relying on theoretical frameworks from the sociology of science. The first one, considered in Chapter 6, is Fuchs' view on intellectual communities and his distinction between specialized and fragmented communities as anchored on the mutual dependence of scholars and their task uncertainty. In the chapter we follow core sources and co-authorships, as well as consider different specialisms in history over time, and explore the question if history is specialized or fragmented and to what extend this aspect is changing over time. The last Chapter 7 gives an explicit operationalization of Becker and Towler's distinction of research specialisms into rural and urban, or intellectually and socially dispersed and concentrated respectively, and applies it to a broader variety of specialisms from history, literature, biology, computer science and astrophysics. The chapter's focus is on the connectivity properties of the bibliographic coupling representations of these specialisms, considering both reference and textual similarities.

Eventually, the main results of the thesis are summarized and discussed in the conclusions (Chapter 8), where their implications as well as perspectives on future work are expanded upon.

2 State of the art

The state of the art presented in this chapter is focused on bibliometric studies of the humanities – history in particular – covering such topics as the characterization of the scholarship produced, its indexation and analysis, as well as the social practices of scholars. We will particularly focus on science mapping studies, as they are methodologically more akin to the approach taken in the present work. We also introduce three theoretical frameworks which will be used in the second part of the thesis and to frame our conclusions. Further discussion and context will be provided at the beginning of each chapter, making reference to this section where appropriate.

2.1 The bibliometric characteristics of the humanities

The amount of scientific publications is growing at an increasingly rapid pace since the 19^{th} century [Bornmann and Mutz, 2015]. In the humanities, this trend has fostered concerns for the apparently overspecialized nature of new research, consequence of the explosion of contributions and novel avenues taken by scholars, all in the absence of developments towards more reliable and effective ways to navigate the existing literature [Tyrrell, 2005]. It would be thus most interesting to understand the process of knowledge accumulation in the humanities, and explore how the current organization and growth of science, coupled with the rising influence of the digital humanities, influences it from a variety of perspectives. Furthermore, the consequent need to advance our understanding in view of 'a bibliometrics for the humanities', must be put in context with a growing demand for the quantitative evaluation of scientific outputs [Hammarfelt, 2016]. In this respect, it does not go without saying that the blind application of metrics developed for the sciences might not be at all appropriated for the humanities.

The humanities are indeed considered to possess a set of characteristics which make it more challenging to acquire and use citation data to study their intellectual organization and communication practices, often preventing the straightforward application of traditional bibliometric reasoning developed studying the sciences (for what follows see *inter alia*: Garfield [1980]; Glänzel and Schoepflin [1999]; Barrett [2005]; van Leeuwen [2006]; Hellqvist [2009]; Linmans [2009]. For reviews see instead: Hicks [1999]; Nederhof [2006]; Huang and Chang [2008]). First of all, humanists often strongly feel the importance of the national and local dimensions in terms of their reading public, the language they use to publish and the object of their research. For a historian, for example, it is impossible to abstract from place and time, which often entail the need to access primary sources which are specifically located and written in local languages. Humanists use a variety of publication typologies and are not just focused on journal articles. It is worth stressing that monographs are especially important, as the practice in the humanities still favors them over other kinds of publications in order to get recognition within the field, despite variations in citation patterns among different disciplines [Knievel and Kellsey, 2005; Williams et al., 2009]. The humanities have also been found to identify core works at a slower pace than other sciences, in part due to longer times required for citation accumulation, entailing a longer life-span of publications [Nederhof, 2006]. The referencing practices of humanists show richer semantic and syntactic usages than in the sciences. For example, recent work has found a significant amount of perfunctory (non-essential) and negative citations in the literature of historians [Lin et al., 2013]. The broad variety of topics and sources being investigated also results in a less focused and wider information retrieval behavior. Manual reference chaining remains important [Buchanan et al., 2005], and browsing is particularly needed for historians, especially so in archives [Talja and Maula, 2003]. The research library remains thus central for locating literature [Stone, 1982]. More recently keyword search, used online and on catalogs or tools such as Google Books, has become more popular [Fry and Talja, 2007]. Humanists indeed increasingly depend now on a myriad of digital tools, which they often use as non-advanced users [Trace and Karadkar, 2017]. Lastly, at the level of their social behavior, humanists have in general little propensity for collaboration and team-work, as attested by co-authorship patterns [Kyvik and Reymert, 2017]. In summary, it is still worth reiterating that 'studies of citation characteristics in the humanities show that the type of publication that is most frequently cited is the monograph, the age span of cited sources is broad, the rate of obsolescence is low, languages other than English play an important role, and self-citations are rare.' [Hammarfelt, 2012a, 34]

Consequently, it is more difficult to build comprehensive citation indexes in the humanities, a condition that hindered bibliometric research in this area [Ardanuy, 2013]. This remains the case despite recent slow progresses, especially made by considering specific fields, important means of publication such as books, and new sources of data [Hammarfelt, 2016]. These considerations in part motivate why, so far, 'the study of the intellectual structure within the humanities using citation analysis is as yet an underdeveloped area' of research [Hammarfelt, 2011].

2.2 Citation indexation

Citation indexation may seem by now a "solved problem" with respect to STEM literature. The crawlers behind mainstream indexes such as Google Scholar (GS), the Web of Science (WoS) and Scopus are largely capable of indexing most citations accurately. To be sure, their coverage is still not complete, albeit improving over time [Mingers and Leydesdorff, 2015; Waltman, 2016; Halevi et al., 2017], both for journals [Mongeon and Paul-Hus, 2016] and monographs [Zuccala et al., 2015]. The main problem which is left open is the skewness in literature coverage and mining performance over different disciplines, with those part of the humanities usually faring worse than most [Harzing and Alakangas, 2016]. Several reasons for this state of affairs have been individuated, which can be grouped into two categories. Intrinsic factors, which depend on the literature and have been discussed above, and extrinsic factors, which depend on the information environment where citation mining is performed, and especially include the variety and fragmentation of supporting catalogs, information systems and other sources of unique identifiers and authoritative metadata. Therefore in the humanities the lack of citation data remains a known problem, lamented several times over [Heinzkill, 1980; Linmans, 2009; Sula and Miller, 2014]. For these and other reasons the use of citations as a means to evaluate research in the humanities has also been questioned [Thelwall and Delgado, 2015; Ochsner et al., 2016a], with alternatives being proposed [Hammarfelt, 2014; Hug et al., 2014; Marchi and Lorenzetti, 2015; Ochsner et al., 2016b; Diaz-Faes and Bordons, 2017; Thelwall, 2017]. In any event, it appears clear that the availability of citation data would not completely solve the issue of research evaluation in the humanities [Hammarfelt, 2017].

As a precondition to citation indexation, the automatic extraction of citations from scholarly publications is a mature area of research. Recent developments include fully fledged architectures to extract and use citation data, embedded within digital library systems [Wu et al., 2014]. Several citation extraction services exist, such as ParsCit [Councill et al., 2008], BILBO [Kim et al., 2011], GROBID [Lopez, 2009], FreeCite[2] and CERMINE [Tkaczyk

[2] http://freecite.library.brown.edu/.

et al., 2015]. In a recent survey and evaluation of several non-commercial reference parsing tools, Tkaczyk et al. [2018] found that the best three performing ones all use Conditional Random Fields (CRF) as the supervised machine learning technique of choice: GROBID, CERMINE and ParsCit, in order. All three benefit from task-specific tuning using extra annotated data, with GROBID showing the best off-the-shelf results. Indeed seven out of the total of thirteen surveyed tools use a CRF approach, while the rest mainly adopt regular expressions. The most recent literature on the topic uses CRF [Körner et al., 2017], Markov logic networks (in a knowledge-based system) [Heckmann et al., 2016] or deep learning [Rodrigues Alves et al., 2018; Prasad et al., 2018]. To date, all published non-commercial reference mining tools rely on these or rule-based methods.

Referencing in the humanities is a less standardized practice than in other sciences, even in the context of the automatic extraction of citations. More specifically, reference lists at the end of a publication are optional, as citations are commonly made in footnotes. Furthermore, humanists developed elaborated practices for the abbreviation and encoding of references, which also entail making a variety of usages of formatting features such as italics or variations in type module. Lastly, it is common in the humanities to refer to both primary and secondary sources. Unfortunately, these characteristics of referencing in the humanities make it difficult to reuse existing services out-of-the-box, as it will become clear in what follows (Chapter 3).

2.3 The core literature and the role of monographs

Given the prevalent importance of journal articles in the sciences, most indexation efforts have not considered citations to books, including scholarly monographs, or are only recently starting to include them. The lack of easily available data from traditional bibliometrics sources, often taken from WoS (Arts & Humanities citation Index, A&HI for short) or Scopus, has thus prevented a thorough investigation of the role of monographs in the humanities. Monographs have been and still are the main publication channel in most disciplines in the humanities [Cullars, 1992; Thompson, 2002; Knievel and Kellsey, 2005; Nederhof, 2006; Larivière et al., 2006a; Williams et al., 2009; Engels et al., 2012; Chi, 2016; Verleysen and Ossenblok, 2017]: 'a monograph may tend to embody a more significant intellectual contribution and a synthesis of a larger body of research than a journal article' [LindholmRomantschuk and Warner, 1996].

As a consequence, the most cited literature in any field in the humanities – its core literature – should essentially comprise monographs [Hicks, 1999], which would benefit in turn from the Matthew effect [Merton, 1968] and become increasingly more popular over time: indeed some studies support this claim. For example, LindholmRomantschuk and Warner [1996] tracked journal article citations to a specific set of monographs, finding a group of core, highly cited sources in every discipline they considered. Hammarfelt [2011] similarly found that 95% of the 200 most cited references were monographs out of a set of journal articles in literature from the A&HI. In a related way, Hammarfelt [2012b] found a well-defined set of core authors in the intellectual base of Swedish literary studies. Furthermore, monographs are also the main cited publication typology of historians [Jones et al., 1972], and are considered to be the most suitable publication typology to be used for bibliometrics studies dealing with the humanities [Chi, 2016]. Nevertheless other studies struggled to find a set of core sources. Such was the case for the seminal work on the history of technology by McCain [1987] or for a study on nineteenth-century British and American literary studies [Thompson, 2002]. If they existed, core works could help improve core collections within research libraries. Yet, in an attempt to pursue this approach, Nolen and Richardson [2016] highlight that 'the combination of these research habits, the diversity of their topics, and the controversial aspect of attempting to define a "core collection" provide very real barriers to identifying, selecting, and acquiring stand-out publications for a library collection in the humanities'. The same authors suggest that a possible reason for this lack of success in identifying core works for humanities' disciplines lies in the focus on the disciplinary macro level, but they are still unable to find core works at the more granular levels of the field or sub-field of study.

A possible motivation for the potential lack of a core literature, or of an easily detectable one, might be that the diversity of citation practices, even within the same discipline, is simply too broad to allow for a set of sources to emerge as a shared core [Thompson, 2002]. Some have instead attributed this apparent absence to the lack of systematic information retrieval practices in the humanities [Barrett, 2005]. It is known that the humanities present great variability in citation practices among their different disciplines [Knievel and Kellsey, 2005]. Indeed Heinzkill [2007] found that over 40% of the monographs cited from a set of articles in English and American literature fall outside of the field as individuated by library classification. The humanities have also been found to undergo an increase in interdisciplinary citing in recent times [Leydesdorff and Salah, 2010; Hammarfelt, 2011; Liu et al., 2017], which is also coupled by a growing international projection [Hicks, 1999; Engels et al., 2012; Gumpenberger et al., 2016; Kulczycki et al., 2018]. This might not help a core literature to emerge: ‘a less demarcated discipline lacking a central core is heavily influenced by other research fields and therefore more interdisciplinary in referencing practices’ [Hammarfelt, 2016]. Furthermore, since publications in the humanities usually accumulate citations at a slower pace [Nederhof, 2006; Linmans, 2009], it follows that it is more difficult to study the most recent literature [Finkenstaedt, 1990], or the impact of individual scholars and institutions [Hammarfelt, 2011]. It appears clear how a thorough exploration of the recent trends in a humanities field and of the role of the core literature should consider citations to monographs either as source or non-source items (i.e. citations from and to, or just to monographs). It must be stressed that a characteristic shared by many studies is the small size of the datasets available, and the difficulty to reach sufficient coverage even within a small field or a specialism.

2.4 On history

For historians too, a set of complementary publication typologies and publication levels exist. Historians privilege the monograph as the key means to publish the final result of a stream of research [Lowe, 2003], which is carried out mostly individually [Knievel and Kellsey, 2005]. Both non-serial and serial publications, despite taking quite some time to age, do become more rapidly obsolescent than primary sources [Jones et al., 1972], that is to say the evidence on which scholars ground their work [Wiberley Jr, 2010]. Primary sources in turn can be subject to transformations which influence their usage patterns, such as indexation and cataloging in archives and libraries, publication in critical editions and digitization. The rapid rise in the number of online available sources, both primary and secondary, might in fact be the most important change in recent historiography, despite the so-far mostly uncritical attitude of historians in this respect [Hitchcock, 2013]. Another under-appreciated source for historians are reference works, such as dictionaries, catalogs and repertories. Eventually, historians are particularly sensible to two forms of localism: linguistic and geographical [Kolasa, 2012], mainly due to the language and location of their sources. To summarize, historians use a wide array of materials, resulting in a likely large fraction of rarely cited items, yet we would expect few of these to be highly cited core sources, with a longer than average life-span, some with an indefinite life span too. In principle and in practice, there is no reason to think that core sources should only be books, even if most likely are.

2.5 Science mapping the humanities

Science can be conceptualized in a variety of ways, for example if can be viewed as a process of accumulation of new knowledge. Maps of science are attempts to localize and relate by relative positioning some entities of interest, such as publications, authors or journals, by way of some relations among them, for example citations [Börner and Scharnhorst, 2009]. Maps of science are especially helpful to uncover the cognitive structure, or intellectual organization of a discipline, in its constituents fields, sub-fields and topics of interest. The scale of analysis, the nature and variety of entities and relations, as well as the dynamics of the scientific process all play a role in this respect. If the mapping of science has been producing a growing number of contributions and increased understanding over time [Börner et al., 2003; Boyack et al., 2005; Börner, 2010; Chen, 2017; Chen and Song, 2017], the situation is less clear for the humanities. Often omitted from maps [Klavans and Boyack, 2009], 'the fine-structures of the humanities have been black-boxed and insufficiently unpacked; the available studies focused mainly on their positions relative to the social and natural sciences' [Leydesdorff et al., 2011].

Usually, the research front can be mapped using bibliographic coupling, its intellectual base using co-citation networks [Persson, 1994]. The core literature, or highly cited sources, plays an important role in the intellectual base: 'the intellectual base is constituted by the core documents of a field; the documents that you should have read or cited, or the "classics" which you at least should be familiar with in order to be recognized as a member of the research community' [Hammarfelt, 2011]. These consideration might not immediately apply to the humanities. The very notion of a research front has in this respect been questioned: 'the being-cited patterns [in some cases] do not indicate the provision of a knowledge base for new knowledge contributions at a research front, but may mean a source of cultural inspiration and influence. This would also explain the slower pace of "progress" in the humanities' [Leydesdorff and Salah, 2010].

Despite the fact that science mapping studies focusing on the humanities are not abundant, we can still find a set of empirical studies which have been carried out. We limit our attention to citation network analyses resulting in mapping efforts, a kind of bibliometric analysis seldom carried out for the humanities in the early decades of the discipline, in favor of more general descriptive citation studies [Hérubel, 1994].

Some studies considered specific journals or disciplines. In an early effort Hérubel and Goedeken [2001] analyzed the French journal *Annales* using the A&HI, and assessing its international reach, as well as its capacity to rely on a broad array of literature from a variety of fields. Leydesdorff and Salah [2010] analyzed two journals in the arts, Leonardo and the Arts Journal, using data from the A&HI. The authors found that both journals cite mostly within the span of their original domain, but are cited widely outside of it. Differently, a small set of articles in the digital humanities was found to cite widely but being cited from a narrower community, resembling the sciences with respect to its 'being-cited patterns'. The authors also add that: 'in the arts and humanities, one focuses on the tips of icebergs of possible references even more so than in the (social) sciences, since publication in the arts and humanities cannot be considered as an endogenous mechanism for generating and supporting a research front.' Coscia and Schich [2011] took the perspective of annotated bibliographies and their classification systems, a regular practice which aims at indexing the all new publications for specific disciplines in the humanities such as classics. By considering the *Archäologische Bibliographie*, a bibliography for classical archaeology from 1956 to 2007, the authors proposed a way to explore both its classification system and, through it, the authors and publications of the bibliography at different levels of scale. They found that 'publications and authors in classical archaeology seem to specialize roughly on certain genres, governed by an either spatial, temporal, or a more generic conceptual perspective.' The literature seems to organize itself by enriching and densifying a skeleton of clusters already in place since the 1950s in the classification co-occurrence network, either signaling that the literature has been incrementally growing in the well-defined fields of classical archaeology, or the conservative nature of the classification system of the *Archäologische Bibliographie.* Weingart [2015] analyzed the fields of History and Philosophy of Science, relying on citation data from the A&HI to both source and non-source items. Using bibliographic coupling and co-citation networks among journals and authors, the author showed how the two communities harbour a third community of authors at their border, who draw from both. Further in philosophy, Ahlgren et al. [2015] explored the 'subdomains' of free will and sorites, using co-citation maps at the level of authors, publications and journals, and terms co-occurrences, using A&HI citation data. Interestingly, the authors found a mapping organized into fields of inquiry for free will, with important connections outside of philosophy proper (e.g. to neuroscience), and organized into smaller topics for sorites, consistent across different networks.

Leydesdorff et al. [2011] provided for the first (and for now unique) time an attempt to map all the humanities using the whole A&HI index for the year 2008. Perhaps the most salient finding was a coherent set of twelve dimensions (latent factors) clearly organized in more or less proximal areas of research, among which we find classics, religion and archaeology; linguistics and the history and philosophy of science; literature and history; arts; music.

Different approaches were also considered. Kreuzman [2001] used author co-citation analysis in the fields of the philosophy of science and epistemology using A&HI data. Multidimensional scaling was used in order to project the co-citation relations on a two-dimensional plane, broadly finding a division of authors according to the field or sub-fields and to the quantitative or qualitative approach. The perspective of author co-citation networks was also taken by Hammarfelt [2012b] for the field of Swedish literary studies, finding a clear set of core, highly cited and influential authors, mainly emerging at an international level or from contemporary Swedish literature.

Yet another different perspective was taken by Zuccala et al. [2015], who ranked scholarly book publishers in historiography using citations to books from articles indexed in Scopus. The resulting map of publishers shows a strong polarity towards prestigious English or American publishers, with only some topical organization. A final aspect of the humanities, which has barely been explored, is mapping the use of primary sources. Romanello [2016] considered data from *L'Année Philologique*, an index of reviews of publications in the domain of Classics. The author was able to make preliminary efforts in the study classical authors, their works and even common quotes through their citation networks.

This overview of mapping efforts in the humanities highlights some commonalities:

- The reliance on existing citation indexes, above all the A&HI, with all its limitations, especially notable its bias for journal articles in English.

- The almost lack of general maps, but instead a narrower focus on specific disciplines or areas of research.

- The presence of several attempts to overcome the lack of data, e.g. by using non-source items, classification systems or novel datasets prepared with considerable effort.

- The substantial lack of explicit connections or inspiration by existing theoretical developments in philosophy or sociology of science focusing on the intellectual organization of the humanities.

Eventually, this overview of the previous bibliometrics literature on the humanities highlights some elements which inform the present work:

1. Citation indexation and data availability: perhaps the single most significant barrier to studying the intellectual organization of any field in the humanities is the lack of systematic data. What is available through commercial citation indexes is heavily skewed towards journal articles in English, thus far from being representative.

2. Intrinsic obstacles to citation indexation and analysis of data: there are intrinsic factors which make it more challenging to perform citation indexation in the humanities, such as complex syntactical referencing practices, and to analyze the resulting data, such as multi-lingual publications and richer referencing semantics.

3. Any analysis of the intellectual organization of a field in the humanities should at least consider scholarly monographs, given their importance in the publication practices of humanists.

2.6 Theoretical frameworks

Research fields can be seen as organizational or epistemological entities, the two views being hardly separable in practice. We briefly introduce here three theoretical frameworks which take different departures in this respect, and will be used both to guide empirical research and to interpret its results in what follows. We focus first on Whitley [2000]'s study on *The social and intellectual organization of the sciences*, which takes an institutionalist perspective that has much influenced subsequent literature. Secondly, we discuss Fuchs [1993a]'s *Sociological theory of scientific change* as an effort to expand upon Whitley's and others contributions and explain scientific change.

Lastly, we discuss the main insights from Becher and Trowler [2001]'s *Academic tribes and territories*, a socio-anthropological study of several scientific fields.

2.6.1 *The social and intellectual organization of the sciences*

Richard Whitley's classic work on the social and intellectual organization of the sciences offers an 'analytical framework for comparing scientific fields as [..] reputational work organizations', where science is 'organized and controlled' [Whitley, 2000, x]. It is by focusing on the variety of ways that this is achieved that 'patterns of intellectual organization' can also be understood. It is worth stressing that the 'basic unit of analysis here, then, is the major organizational entity controlling access to jobs, facilities, technicians, and other resources through public reputations' [Whitley, 2000, 164]. Whitley's framework has been previously used to interpret empirical results in bibliometrics. Most recently, Bonaccorsi et al. [2017] found a macro similarity at the level of indicators of performance across disciplines (log-normal), despite substantial micro variations, suggesting to use the reputational framework as an explanatory perspective.

The proposed analytical framework uses two main axes: the mutual dependency among scholars and their degree of task uncertainty. By *mutual dependency* it is meant the degree in which a scholar depends on others for conducting his research. Whitley distinguishes between a *functional dependency*, which focuses on the direct reliance of researchers on the results of colleagues in order to advance the field, and *strategic dependency*, which is instead more broadly defined as how important it is for researchers to agree on results for the field as a whole. Strategic dependency has ultimately to do with the importance of agreement and the criteria for acceptance shared within the field. The degree of *task uncertainty* relates instead more directly to the intellectual organization of the field, and has to do with the predictability of results as well as questions posed. Task uncertainty can be *technical*, if it deals with the interpretation of results and the choice and application of methods, or *strategic*, if instead it is concerned with the broader goals of research and choice of relevant problems for the field. High task uncertainty means there is much debate around results and methods while high strategic task uncertainty entails there is a low consensus among scholars on what are the shared goals of the field.

Whitley's suggests then a classification of scientific fields according to how they are institutionally positioned with respect of these axes, as follows [Whitley, 2000, 158, Table 5.2]:

- *Fragmented adhocracy*: 'producing diffuse, discursive knowledge of commonsense objects'. Characteristics: low functional and strategic dependence, high technical and strategic task uncertainty. Examples given: political and literary studies, British sociology.

- *Polycentric oligarchy*: 'producing diffuse, locally co-ordinated knowledge'. Characteristics: low functional and high strategic dependence, high technical and strategic task uncertainty. Examples given: German philosophy, British social anthropology.

- *Unstable*. Characteristics: low functional and strategic dependence, high technical and low strategic task uncertainty. No examples given.

- *Partitioned bureaucracy*: 'producing both analytical specific knowledge and ambiguous, empirical knowledge'. Characteristics: low functional and high strategic dependence, high technical and low strategic task uncertainty. Examples given: Anglo-Saxon economics.

- *Professional adhocracy*: 'producing empirical, specific knowledge'. Characteristics: high functional and low strategic dependence, low technical and high strategic task uncertainty. Examples given: bio-medical science, artificial intelligence, engineering.

- *Polycentric profession*: 'producing specific, theoretically co-ordinated knowledge'. Characteristics: high functional and strategic dependence, low technical and high strategic task uncertainty. Examples given: continental mathematics.

- *Technologically integrated bureaucracy*: 'producing empirical, specific knowledge'. Characteristics: high functional and low strategic dependence, low technical and strategic task uncertainty. Examples given: chemistry.

- *Conceptually integrated bureaucracy*: 'producing specific, theoretically oriented knowledge'. Characteristics: high functional and strategic dependence, low technical and strategic task uncertainty. Examples given: physics.

Table 1: Characteristics of the internal structure of seven major types of scientific field [Whitley, 2000, 169, Table 5.3].

	Configuration of tasks and problem areas			
Types of scientific field - Feature	**Specialisation and standardisation of tasks and materials**	**Degree of segmentation**	**Degree of differentiation into schools**	**Hierarchization of sub-units**
Fragmented adhocracy	Low	Low	Low	Low
Polycentric oligarchy	Low	Low	High	Low
Partitioned bureaucracy	High in core, medium in periphery	Medium	Low	High
Professional adhocracy	High	Medium	Low	Low
Polycentric profession	High	Medium	High	Low
Technologically integrated bureaucracy	High	High	Low	Low
Conceptually integrated bureaucracy	High	High	Low	High
	Co-ordination and control processes			
	Impersonality and formality of control procedures	**Degree of theoretical co-ordination**	**Scope of conflict**	**Intensity of conflict**
Fragmented adhocracy	Low	Low	High	Low
Polycentric oligarchy	Low	High	High	High
Partitioned bureaucracy	High in core, medium in periphery	High	Low	Medium
Professional adhocracy	High	Low	Medium	Low
Polycentric profession	High	High	Medium	High
Technologically integrated bureaucracy	High	Medium	Low	Low
Conceptually integrated bureaucracy	High	High	Low	Medium

The likely candidate for most disciplines in the humanities, including history, would be fragmented adhocracies. These are characterized by a 'rather personal, idiosyncratic, and only weakly co-ordinated' research, with a clear connection to the general, educated public and therefore a sometimes blurred boundaries between professionals and amateurs. In fragmented adhocracies 'commonsense languages dominate the communication system' [Whitley, 2000, 159], in a fluidity of standards and substantial openness of the reputational system. Given the assignment of a certain research field within a category, the specificities of its social and institutional organization will result, for Whitley, in 'different patterns of intellectual organization. The location of a particular field in one type of science, [..] implies a certain way of structuring research and a certain characterization of its knowledge' [Whitley, 2000, 165]. A set of features are then proposed in order to characterize intellectual configurations resulting from different institutional organizations. The two main groups of features are the configuration of tasks and problems within the field, and 'the means by which, and degree to which research is co-ordinated and controlled across research sites and groups' [Whitley, 2000, 166]. The resulting characterization is provided in Table 1.

In this configuration, fragmented adhocracies are characterized by 'intellectual variety and fluidity'. An important aspect of their organization is the premium given to originality, the idiosyncratic methods and individual research strategies, the lack of efforts to integrate results, so much so that the differentiation of contributions is a higher priority here than co-ordination of results and contribution to the collective enterprise' [Whitley, 2000, 174]. In summary, given the absence of much theoretical work and that specialization is seen as a way to accomplish differentiation and avoid integration of results, 'intellectual fragmentation is [..] a dominant feature' [Whitley, 2000, 176].

Whitley's framework, despite its interest, might be perhaps too restrictive in its determination of intellectual organizations resulting from institutional configurations. Indeed, within an institutionalized discipline, a variety of possibilities might co-exist. For history in particular, as it will become clear in due course, not all knowledge is similar and significant differences might exist between, for example, work in so-called ancillary disciplines (archival studies, palæography, diplomatics) and historiography proper, despite identical institutional affiliations. In conclusion, Whitley's analytical framework might convey a significant explanatory power with respect to the institutional organization of science, it remains to be seen to what extent the intellectual organization reflects institutional boundaries, so much so in periods of substantial innovation as with the digital humanities turn.

2.6.2 *A sociological theory of scientific change*

A clear sociological and organizational underpinning is also informing Fuchs [1993a]'s theory of scientific change, part of his broader theory of scientific organizations (TSO) [Fuchs, 1992]. The TSO views research specializations as 'reputational work organizations', exactly as in Whitley's. There are three aspects to the TSO: a sociological perspective stating that the ways we think and perceive the world are shaped by social structures (e.g. mutual dependence among scholars); that social and cognitive structures are also informed by the way work is done and technology is used (e.g. task uncertainty and the role of research technologies); and a materialist theory of consciousness by which the ways we think are related to the control of the 'material means of mental production' (e.g. the centralization of technologies in Big Science projects). A central tenet of the theory are the (recurring) axes of mutual dependence and task uncertainty, borrowed from Collins [1975] and Whitley [1984]. Fuchs' most significant contribution is instead a focus on scientific change and its explanation through *competition*.

Fuchs argues that, within reputational work organizations such as academic research fields, competition is a major drive of change, and the way competition acts vary according to the two axes of mutual dependence and task uncertainty, as shown in Figure 1. The three main effects of competition are: to foster a state of permanent discovery in fields which possess both high mutual dependence and task uncertainty; to instead push towards increasing specialization and knowledge cumulation in fields with high mutual dependence but low and predictable task uncertainty; finally, to generate fragmentation in fields with low mutual dependence but high task uncertainty, such as often in the humanities. Eventually, fields in a regime of both low task uncertainty and mutual dependence are seen as stagnant. An important insight is that different ways of doing science actually coexist and are practices by different cohorts of researchers within fields; most notably a group of leaders can proceed for permanent discovery at the core while many more researchers work cumulatively at the periphery.

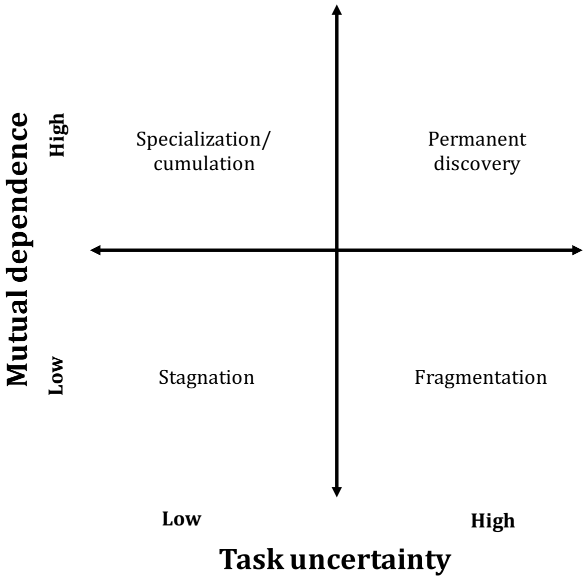


Figure 1: Three types of scientific change [Fuchs, 1993a, 940, Figure 1].

Most relevant for our purposes is the distinction between specialization and fragmentation. Essentially, both are effects of similar strategies to cope with competition and, it will be argued, information load. The different outcomes relate to the presence or absence of a 'paradigmatic integrity in the larger field' which can be used to effectively integrate new contributions in the larger body of knowledge mostly via a shared theoretical framework. Fragmentation ensues from similar process but in the absence of such framework. The humanities, and history too, are notoriously theory-adverse, yet integration might come from other sources, as we will discuss in the second part of this work.

2.6.3 *Academic tribes and territories*

Becher and Trowler [2001] offer a broad study of a variety of fields of research and education, taking an ethnographic approach by using interviews. Their framework was already put forward in Becher [1989]. Since this study is somewhat bottom-up, their framework reflects an attempt to accommodate a variety of perspectives using simple yet intuitive concepts and metaphors, and not as much the superimposition of a theory on empirical evidence. Their main distinctions are between fields which: produce soft/hard and pure/applied research, organize intellectually in rural or urban landscapes, and are socially convergent or divergent. The rural/urban metaphor relates to the intellectual organization of a field into many, small topics of investigation, where research is loosely coordinated and proceeding at a slow pace (rural, the humanities), or in few, larger topics with many researchers active in each of them, producing research at a fast pace (urban, many sciences). We amply discuss and use this conceptualization in Chapter 7.

The distinction between soft or hard mainly relates to the degree of abstraction or concreteness the produced knowledge possesses, with reference to the *nomothetic* and *idiographic* distinction. Pure or applied instead reflects the influence a field receives from the outside, or the use of results: if they are limited to other sciences (pure) or instead to the outside of academia (applied). A typical hard-pure science is physics, where knowledge is cumulative and concerned with universals, quantitative and oriented to produce explanations, consensus is high and there are clear criteria for knowledge verification. The humanities belong to the soft-pure category, as they produce holistic knowledge through reiterations, deal with particulars and their understanding, lacking shared criteria for consensus. Technical fields are hard-applied, including medicine, engineering and the like. They produce pragmatic and purposive knowledge, often resulting in products, methods and technologies. Lastly, in the soft-applied category we may find the social sciences, such as law or education, producing knowledge in the form of procedures or protocols, of a functional or utilitarian scope [Becher and Trowler, 2001, 36]. Evidently, some fields elude clear categorization, such as economics which might be seen as hard-soft-applied.

If the categories of soft, hard, pure or applied mainly refer to the nature of knowledge which is produced, and its use, the convergent and divergent categories reflect instead social configurations. A convergent research field possesses an elite of scholars which guide and control it, acting as brokers, hubs and gatekeepers and giving a socially cohesive aspect to it. Divergent fields lack this elite, or possess a multiplicity of independent or even conflicting elites [Becher and Trowler, 2001, 184]. Physics is mainly convergent, for example, while most humanities are divergent. Another aspect touched upon is the use of language and the communication practices of different fields [Becher and Trowler, 2001, Ch. 6]. Language can be specialized or everyday, with a level of technicality largely depending on the target audience and the need for lack of ambiguity. Some fields also put a premium on a thorough and self-contained explanation of their reasoning, which in part explains the preference for books in fields with slow literature ageing such as the humanities, while others instead favor short, clear-cut, to-the-point even if stylistically dry publications. The aspect of language will be further expanded in Chapter 7 too.

To conclude, we saw that these three theoretical frameworks share much similarities despite taking different starting points. Whitley's view of research fields as reputational organizations stems from an institutional focus which has in turn consequences on their intellectual organization. Fuchs instead, by bringing forward competition and coordination, puts the two aspects of the social and intellectual organizations on a more even footing. Becker and Trowler's propose a less coherent framework where different metaphors are used in a loosely coordinated way to describe different realities of both the social and intellectual organization of research fields. Their perspective is anthropological and sociological at the same time. These frameworks offer by far and large similar views on research fields. In what follows, we will consider Fuchs' distinction between specialization and fragmentation in Chapter 6 and Becker and Trowler's metaphor of rural and urban fields to compare the humanities and the sciences in Chapter 7. We will then rely on Whitley's theory in order to interpret our results in the conclusions. Since there is no clear-cut boundary to be drawn between the social and intellectual organization of research fields, any investigation of the intellectual aspects must be eventually informed by the social context where its scientific body of knowledge was produced.

3 Citation indexing in the arts and humanities: the case of the historiography on Venice

The scholar looking to explore the intellectual organization of any field of research in the arts and humanities faces a hard-to-avoid obstacle: the fragmentary and all too partial coverage of data sources, or lack thereof (see Section 2.2, for what follows in this introduction). Both citation and publication corpora suffer from the same problem. Scholarly publications in the arts and humanities are considerably less impacted by open access or pre-print [Piwowar et al., 2018], often keep being published primarily or exclusively on paper, and their digitization and indexation backlog is proportionally larger than for the sciences. Consequently, citation data is largely missing. At best, we have at our disposal data from citation indexes thought from and for the sciences, such as the Web of Science or Scopus, where the journal article is the focus, a form of publication likely of secondary importance in the arts and humanities.

In this chapter we suggest that three properties of publication and citation data sources are important to study the intellectual organization of research fields in the arts and humanities in their richness: I) their coverage in terms of *publication typologies* (especially including books) and *cited sources* (to consider primary sources). Coverage should be as wide as possible, and at least representative of all main publication and cited source typologies. II) Their *chronological depth*, since the process of knowledge accumulation is slower than in most sciences, and the scholarly traditions in the arts and humanities are likely enduring for a possibly much longer time. Ideally, the chronological depth should span to the birth of modern academia during the 19^{th} century, possibly more. III) The *sampling strategy*: since the publication venue (e.g. the journal) or language or publisher cannot alone yield a representative sample of the literature in a field in the arts and humanities, other external factors should be explored in this respect. In particular, the collections of research libraries might offer a better means to individuate representative corpora.

In order to produce a publication and citation dataset following the abovementioned, quite demanding requirements, a single and circumscribed case-study is considered: the modern historiography on Venice. This choice was made for three reasons: the field is medium sized, with an amount of scholarly literature in the order of thousands of volumes; the field also has an intellectual history influenced by the encounters and exchanges of the local community of scholars and several international ones, which makes for a particularly compelling case of study; lastly, the presence of parallel digitization projects conducted by the École Polytechnique Fédérale de Lausanne and the University of Venice taking place at the Archive of Venice allowed to develop strong synergies with the digitization and indexation effort conducted on the secondary literature. A sizeable amount of scholarly literature on the topic has been digitized and the citation therein extracted, allowing to study the intellectual organization of a community of historians at an unprecedented level of detail. The dataset on Venice enabled in part to complement the use of traditional bibliometric data sources in the present study. Not least, this work fostered the development of a more general approach to the indexation of the literature in the arts and humanities, illustrated here by a software prototype.

This chapter is organized in the following way: we start by giving an overview of the approach developed in order to digitize and index a corpus of publications on the historiography on Venice, respectful of the three properties discussed above (Section 3.1). We then delve into testing the guiding assumption of this approach, namely the reliance on library collections to produce a citation dataset representative of a domain of interest (Section 3.2). Eventually, in Section 3.3 we generalize our work by proposing an approach for the collaborative and open citation indexation of the literature in the arts and humanities, which stems from the experience with the historiography on Venice.

We borrow from the following published or forthcoming contributions: Colavizza and Kaplan [2015] delves into the problem of parsing references to primary sources; Colavizza et al. [2017] considers the problem of how to span the literature of a field using library resources and an iterative approach, mainly relying on scholarly monographs and reference works (Section 3.2); Colavizza and Romanello [2017] is a dataset release of manually annotated references from the literature on the history of Venice; Rodrigues Alves et al. [2018] provides a deep learning architecture for reference parsing in the arts and humanities and Colavizza et al. [2018c] sketches the general idea for a collaborative citation index in the arts and humanities (Section 3.3). The open code and data resulting from work discussed in this chapter is provided in Chapter 9.

3.1 Overview of the approach

The construction of a citation index for the historiography on Venice had to start essentially from scratch, since no digitization was already accomplished. This challenge provided for the opportunity to start from little assumptions, with the important question of what constitutes the literature on the history of Venice, and how to find it. The approach taken can be organized into the following steps, as illustrated in Figure 2:

- *Corpus selection and acquisition*: in this step a corpus of publications is selected and digitized. The selection of the corpus can rely on a variety of external resources, such as domain experts, library catalogs and collections. Furthermore, it can benefit from an iterative approach, where a seed of recent publications is digitized first, in order to progressively discover new literature via their citations. The acquisition is instead an essentially technical step, whose main conceptual challenge is the problem of copyrights.

- *Target analysis and annotation*: once a corpus is acquired, including the full-text of publications, the following step entails preparing the ground for the automated extraction of references. In particular, this requires a preliminary study on referencing styles and practices present in the corpus, and the manual annotation of a sufficient amount of references in order to train a machine learning method to extract them from the rest of the literature under consideration.

- *Reference mining: detection, extraction and classification.* The subsequent step is the task of reference mining, entailing the automatic detection, extraction and classification of references found anywhere in the corpus, including footnotes. Typically, supervised machine learning techniques are used at this stage.

- *Disambiguation of references into citations.* Once references are extracted, they need to be disambiguated, i.e. linked with a unique identifier ideally from an external authoritative source, such as library catalogs. In this way citations, or links between two publications, are established.

- *Publication and (re-)use of project outputs.* Eventually, the publication and citation data can be published and exposed via a variety of services, including a citation index, and used for further analysis.

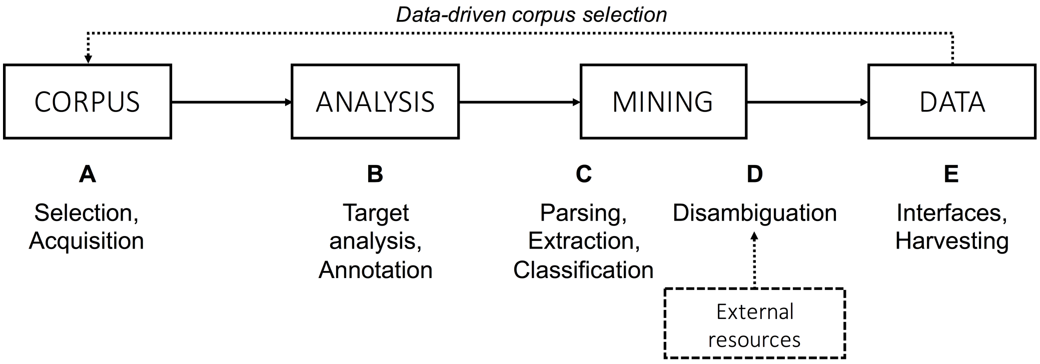


Figure 2: The proposed approach and pipeline to create a citation dataset for the literature on the history of Venice, or more generally any field in the arts and humanities.

The following section further elaborates on this overview, discussing the specific choices made in the context of the present work.

3.1.1 Corpus selection and acquisition

The historiography on a topic, such as the history of Venice, is not easy to bound. A variety of contributions are published, with no existing citation index coming even close to adequately represent the field. As a consequence, we consider our corpus as unbounded from the very beginning, and progressively and gradually enlarge it by using citation data to discover new items for indexation. A first selection of the literature is made using a combination of library catalog look-up (searching by subject and keyword), domain expert advice (mainly for journal titles), existing published bibliographies, user preferences as mapped by rapid-access shelving strategies in specialized research libraries in Venice itself. This selection comprises an initial corpus of circa 2000 books and 10 local journals mostly in Italian (552 issues, for 5496 individual articles), which is used for a first digitization campaign. While the selection of journals is straightforward, as we cover all issues of a given title, that of books is less obvious. The use of library resources entails that we focus on a corpus of relatively well-known or recent scholarly monographs and works of reference. The extent to which this choice is "bibliometrically sound", thus it allows to span the relevant literature on Venice via its citations, needs to be assessed.

Once the corpus has been selected, it can be digitized and integrated with its catalog metadata. Subsequently, the images are OCRed (Optical Character Recognition), using a commercial solution tuned for the specific materials at hand. Some further decisions are made as follows:

1. Copyright: we establish partnership with libraries in possession of the required literature, which is made available under agreement that a digital copy is to be given back to the holding library, that only temporary copies could be used for the purpose of reference mining, and that references and citations, once extracted, would not per se constitute a violation of the publisher's or author's copyright, not constituting an integral part of the contents.

2. Metadata: we acquire a copy of the whole Italian National Catalog, correcting or complementing it as necessary. For example, an important project-specific information is the library provenance of the item, for copyright and source verification reasons.

3. Digitization and OCR: we can make some technical choices in view of subsequent needs. For example, during OCR we also extract layout features (such as font size and usage of italics or bold, often used in footnote references), in order to use this information for reference mining.

The resulting output is a collection of paired images and text files, one for each page of a bounded volume (book or journal issue). More details are given in Section 3.2.

3.1.2 Target analysis and annotation

Our aim is to index all cited sources, in all their forms. The main general typologies of cited sources, with respect to the structure of their references, are:

1. *Primary sources*: any documentary evidence, either in original or edited, or non-scholarly publications.

2. *Secondary sources, books* including monographs.

3. *Secondary sources, articles and contributions*: any publication contained within another publication, such as edited volume, journal issue, and the like. In this case, references contain both the details of the specific publication and of its container publication.

Furthermore, references can be given in full, or in a variety of abbreviated forms, highly dependent on context. For example, the full primary source reference:

`"Archive of Venice, Procuratori di San Marco, de citra, commissarie, b. 1, c. 7."`
Components: archive, record group, series, sub-series, box, sheet.

Can be abbreviated as:

A: `"ASVe, PSM, de citra, commissarie, b. 1, c. 7."`

With acronyms defined at the beginning of the publication, or even:

B: `"Ivi, c. 8."`

To refer to another sheet (c. 8) in the very same box as in the immediately previous reference. The procedure is similar for any kind of cited source. We identify A as a *global abbreviation* (dependent on the global reference context of the publication) and B as a *local abbreviation* (dependent on the local reference context, i.e. previous references). Furthermore, given the variety of the literature, it is profitable to conduct a reference practice survey to inform the techniques to adopt in subsequent steps.

A preliminary and exploratory annotation campaign is subsequently conducted, focusing on the broadest variety of publications possible. This campaign allows to establish a classification of cited sources and their possible abbreviations, and a classification taxonomy for reference components (author, title, year, archive, etc.). This campaign goes on until a) major changes to the taxonomy no longer occur; b) every element in the taxonomy is reasonably represented in terms of the frequency of its occurrence. Note that an estimate of the relative importance of each element in the taxonomy is relevant for its consolidation during reference mining. Afterwards, all these preliminary annotations are discarded. The details of the resulting taxonomy are given in Colavizza and Romanello [2017].

3.1.3 Reference mining: detection, extraction and classification

Given the availability of annotated data, and having reached a stability of the annotation taxonomy, we use supervised learning methods for detecting, extracting and classifying references, experimenting with both Conditional Random Fields [Colavizza et al., 2017] and deep learning [Rodrigues Alves et al., 2018]. We frame the tasks as follows:

1. A first parser, considering the full text of every publication in the collection, tags individual tokens with specific tags (such as author, title, year of publication). The unbalanced quantity of annotations for more rare tags requires some consolidation of the taxonomy, to group rare and similar tags together.

2. A second parser, using the output of the first one, tags every token in the text as being outside, inside, beginning or ending a reference, plus assigning the general typology to it (primary source, secondary source - book or secondary source - journal article). For example, the following footnote (number 5):

   `"(5) A.S.V., Provveditori sopra monasteri, b. 280; Riformatori dello Studio di Padova, f. 272."`

   Is parsed at first yielding the following result:

   `"(5)"` out of reference.
   `"A.S.V.,"` archive.
   `"Provveditori sopra monasteri,"` record group.
   `"b. 280;"` box.
   `"Riformatori dello Studio di Padova,"` record group.
   `"f. 272."` sheet.

   Then parsed a second time yielding the following result:

   `"(5)"` out of reference.
   `"A.S.V., Provveditori sopra monasteri, b. 280;"` primary source.
   `"Riformatori dello Studio di Padova, f. 272."` primary source.

The (ideal) end result is thus the extraction of two references, with their components and general categories.

An explicit choice we make at this stage is to maximize the recall evaluation score, at the expense of precision.[3] This approach yields many false positives, or tokens tagged with specific tags and general typologies, despite not being part of a reference. Examples are book or article titles or in-text mentions of authors and other named entities. Nevertheless, as we will discuss in what follows, using high recall at this stage and high precision for the subsequent disambiguation task, results in a balanced pipeline at the end.

3.1.4 Disambiguation of references into citations

Having mined references, the following task is to disambiguate authors and cited sources. We rely on a set of resources to link references to their referred items. First of all, most citations to books can be matched using the Italian National catalog[4], which we deploy locally. Secondly, we rely on the information system of the Archive of Venice, called SiASVe[5], which was likewise replicated locally, and where every record group and document series of this archive is named, indexed and described. Lastly, for authors we use VIAF (Virtual International Authority File), which is an OCLC publicly available author authority record system[6]. Despite our best efforts, several references to items not to be found in these systems, such as other primary sources or journal articles, are left out.

For these reasons, the disambiguation task is approached as follows:

1. A first internal search is performed on already disambiguated items. If a match is found, the process stops.

2. If no match is found, a search is conducted on the three systems, if the general typology of the reference is appropriate. If a match is found, the system stops.

3. If no match is found, a new entry in the index is added.

[3] In a binary classification task there can be four possible outcomes: a true positive is a correct positive classification (TP), a true negative is a correct negative classification (TN), a false positive is an incorrect positive classification (FP) and a false negative is an incorrect negative one (FN). We define precision as $p = \frac{TP}{TP+FP}$ and recall as $r = \frac{TP}{TP+FN}$.

[4] `http://www.iccu.sbn.it`.

[5] `http://www.archiviodistatovenezia.it/siasve/cgi-bin/pagina.pl`.

[6] `http://viaf.org`.

Searches are also performed with different methods according to the external resource under consideration. The internal lookup, the Italian Catalog lookup and the VIAF lookup use a combination of string and rule matching. The SiASVe lookup uses a supervised multinomial logistic classifier, rolling back to rule matching if the probability for a classification is low (and therefore the referred item is likely not part of the training data). It is worth noting that this classifier relies on a large amount of manually corrected disambiguations.

The disambiguation step is presently still far from reaching satisfying and robust results. The frequent lack of either external authority systems, or the lack of APIs (Application Programming Interfaces) to access them, greatly hinders our efforts to rely on external resources. At the same time, the reliance on internal lookup is limited by the quality of reference data.

3.1.5 Publication and (re-)use of project outputs

The results of the indexation of the scholarship on the history of Venice are already in part released: the annotated references for reference mining along with code implementing the machine learning parsers, and a first citation dataset of book citations which will be considered in what follows. Two interfaces, a digital library and a citation index, are also published to make the corpus accessible. Lastly, the code developed during the project, as well as an API giving full access to the citation data with an associated SPARQL endpoint,[7] are both planned as future developments.

The work of creating a citation dataset from scratch proved more challenging and involved than anticipated, yet it fostered an effort of generalization which we also discuss in what follows. Its primary purpose, that of creating datasets for the in-depth analysis of a field in history, is only partially accomplished to this date, yet we could already complement traditional bibliometric data sources in the analytical effort which is the object of the present study.

In the following section we consider part of the corpus on Venice, namely books, and discuss in detail the application of this approach to it. Notably, we assess to what extent the field of the history of Venice can be covered by using references extracted from books part of a research library collection.

[7] A query language for semantic web databases.

3.2 The references of references: creating a new citation corpus

The work of humanists does not fully rely on citation indexes, but instead requires to collect sources using a variety of means, including laborious manual citation chaining. Mainly for this reason, research libraries in the humanities play a pivotal role, as they are often able to build over time collections responding to most scholarly needs [Kellsey and Knievel, 2012]. One way research libraries help scholars is by devoting entire physical sections to *reference works and monographs* in specific fields. These references are deemed of importance within a domain of study. Their selection is usually done by librarians and domain experts, for the purpose of accelerating the retrieval of relevant literature or information by users. We argue that library resources can be used to find a corpus of literature relevant in a field of study, and furthermore that reference monographs in particular can play a key role to find new literature to index through their incoming and outgoing citations. We therefore propose an iterative method to create a citation index for the literature in any field in the humanities relying on the *references of references.*

Let us start by introducing some definitions:

- *Reference works*: the part of the literature in a research library identifiably related to a specific domain, which contains scholarly contributions in support to other scholars. Reference works include, but are not limited to catalogs and inventories, historical dictionaries, repertories, bibliographies. In general, reference works are scholarly efforts to list and organise, prepare, distil, summarize or otherwise convey primary evidence or relevant information to other scholars, to facilitate and accelerate their work.

- *Reference monographs*: the part of the literature in a research library identifiably related to a specific domain, which contains scholarly contributions in prose. Reference monographs are the most accomplished expression of the interpretive work of historians.

- *Core literature*: the part of the literature which is considered foundational for a given domain of study, and as such it is highly cited by the rest of the literature. The core literature can be of any of the above kind.

Assuming the existence of a set of reference works and monographs (references for short) for a given field in the humanities, we want to use it as a first seed to extract citation data. References can be identified by the means of a research library: their physical location (e.g. rapid consultation shelves), catalog subject headings, scholarly bibliographies. These are, in practice, the only methods available to scholars to find relevant literature, excluding the process of reference chaining and domain or tacit knowledge. This approach should therefore help focus our efforts towards a subset of the literature which is likely to be of interest for the community at the present time. The main open question by taking this approach is to what extent these references can span the literature of the domain of interest.

We detail here on the application of the approach discussed in Section 3.1 to the reference works and monographs on the history of Venice at the Humanities Library of the University of Venice. Besides providing technical details and producing a dataset of book to book citations which will be analyzed in depth in subsequent chapters, we also set out to test the approach. In particular, we will assess to what extent the citations extracted from these references span within or outside of the initial corpus, and if the of core literature of the field emerges as a result. In what follows we leave out journal articles, to focus instead on publications in the form of books. We also focus on books with reference lists at the end, in order to get better reference extraction results, and therefore in particular on reference monographs. The dataset and code to replicate our results are publicly available, see Section 9.

3.2.1 Approach

Corpus selection, acquisition, target analysis and annotation Libraries can provide a first means to identify a set of references of interest within a specific domain of study, even more so if they specialize in this domain. In our case, the Italian library catalog is used in order to extract:

1. all the resources on reference shelves marked as “History of Venice”, from now on defined as (rapid) consultation works;
2. all the resources under subject history of Venice (e.g. Dewey code 945.31);

3. additional resources found by keyword search over the title (e.g. using words such as “Venice” in multiple languages) and manually selecting what is relevant, or by means if scholarly bibliographies (e.g. Zordan [1998]’s repertory).

The outcome is a set of 1904 books. Within these, we individuate 836 reference monographs with a list of references (of which 201 are in rapid consultation), or 44% of the total. Such monographs with reference lists are also equally distributed over time, as shown in Figure 3. Of these, 700 (183 in rapid consultation) have structured lists of references, as opposed to end notes. This last subset of 700 monographs with structured reference lists is used to extract citations. The distribution of the number of references made by these 700 monographs is given in Figure 4. Values are reasonably between 20-30 and 300; more extensive reference lists are rare. Ideally, we should consider the whole corpus, yet extracting references from footnotes is considerably more challenging than just using reference lists. At the same time, we verified the absence of a systematic bias with respect to topic) as mapped by library classification), period or publisher influencing the presence or absence of reference lists, allowing us to proceed with this subset of the original corpus.

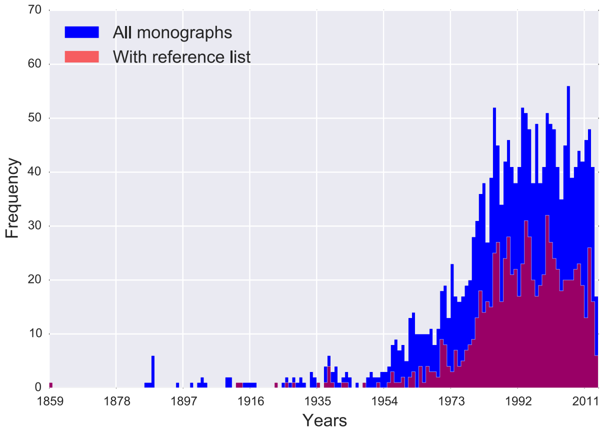


Figure 3: Number of works in the corpus per year (blue/grey), over the monographs with a reference list (red/black): reference lists are uniformly distributed over time with respect to the distribution of the full corpus.

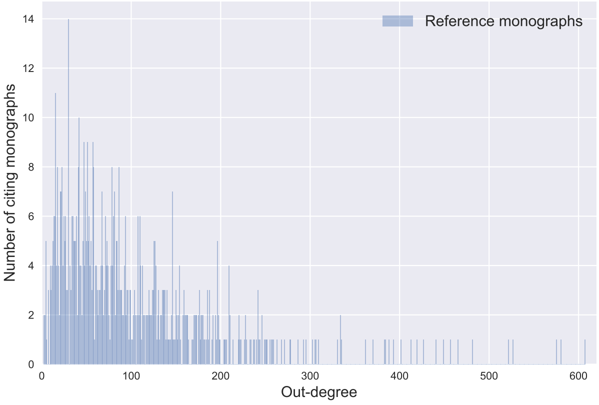


Figure 4: Histogram of the number of references made by the 700 citing monographs.

The second step of the approach is the study of reference styles and the manual annotation of a subset of references for each individuated class. Reference styles can be quite varied in the humanities, and change over time, author and publisher. Yet they convey important information for reference extraction. A *reference style* is a specific combination of elements in a reference, such as author and title, encoded in a predefined way (e.g. using quotation marks for the title). Styles can be grouped in *classes* and *families*. For example:

> “De Virine, Theodore Low. Notable Printers of Italy during the Fifteenth Century. New York: The Grolier Club, 1910.”

is a reference presenting the author’s surname, then name separated by comma, title, place of publication, publisher and date. The punctuation and capital letters in use are particularly relevant. A different class stems from the elimination of at maximum one element, or one change in encoding. E.g. removing the publisher would create a new class of the same family. A different family is identified by at least two removals or additions of elements, and/or sensible changes in the encoding of the same information. For example:

> “De Virine, T. L. Notable Printers of Italy during the Fifteenth Century. The Grolier Club, 1910.”

would create a different class in a separate family as the author's name is now abbreviated and the publisher has been dropped. Classes and their families used as a feature for parsing, have improved results in a sensible way, since the citations from a specific publication all belong to a unique class/family combination. In total we individuate 33 classes and 6 families.

Manual annotation is done over a randomly selected subset of citations for each class.[8] Annotations are divided into two categories: *generic* and *specific*. A generic annotation distinguishes the completeness of a reference (if full or abbreviated) and the type of object referred to (if a book or a contribution, such as a journal article). There can be 6 generic types of annotation: book, contribution, or article, all either full or partial. Specific annotations identify instead the components of generic categories. Examples of specific annotation tags are: "author", "title", "publisher".

Approximately 27% of the 700 monographs have been annotated, 2 pages of references each on average. As a consequence, circa 3.8% of all available pages with references have been annotated, for a total of 49'580 annotations, of which 8646 are generic (i.e. full references) and 40'934 specific (i.e. their components).

Reference mining: detection, extraction and classification The next component of the pipeline is the reference mining module, which performs two tasks:

1. *Reference parsing*: given a text stream of lists of references, parse the text to assign the most likely specific tag to each token.

2. *Reference extraction and classification*: given a stream of tokens with specific tags, decide where a reference begins and ends, and assign a generic category to the reference ("monograph", "abbreviated reference" and "contribution").

Both parsers use Conditional Random Fields with the same set of features—except for specific tags resulting from task 1 that are used in task 2—a technique commonly adopted in similar settings, introduced by Lafferty et al. [2001]. The order of the tasks has been determined empirically to maximize performance on a subset of specific tags (crucially author, title and year of

[8] Using the Brat annotation environment available at `http://brat.nlplab.org` [Stenetorp et al., 2012].

publication), which are the most relevant for the look-up module. 8051 annotated references are used for training and testing, for a total of 122'612 tokens, or circa 15 tokens per reference, plus 35'124 negative tokens (outside of references).

Disambiguation of references into citations via catalog look-up Extracted references need to be disambiguated to establish citations. This task is performed by a look-up system that attempts to match the components of the extracted reference against a library catalog. Given the kind of data at hand, such look-up system needs to: have a good coverage of the domain; have the ability to work with a limited set of metadata fields as input; be robust to OCR errors.

The implemented solution attempts to match the metadata fields of the extracted citation against the bibliographic records contained in the catalog of the Italian National Library Service (SBN), which at the time of writing contains almost 16 million entries. This catalog provides a good coverage of the publications cited by reference monographs, which can be easily explained in light of the focus on the history of Venice. A full dump of the SBN catalog is used, thanks to an ongoing collaboration with the Central Institute for the Union Catalog of Italian Libraries and Bibliographic Information (ICCU), which owns the SBN catalog, and is responsible for its maintenance and updates.

The data are loaded onto an instance of ElasticSearch[9] as it constitutes an efficient solution to search through such a large dataset. The catalog dump is in a JSON format derived from MARC21, the format in which the catalog records are originally stored. Each publication in the catalog is described according to the Italian national guidelines.[10]

3.2.2 Results

Reference mining The reference mining module comprises two supervised models. The first one performs the following: given a stream of text likely to contain a list of references, it initially tags every token with specific tags. A second model then parses the text again in order to attribute generic and begin-end tags at the same time. Eventually, all individuated references for

[9] A full-text search engine based on Apache Lucene.

[10] The last version is detailed in ICCU [2016].

each monograph are exported to the look-up module. The implementation is done in Python, relying on the CRFSuite [Okazaki, 2007]. The set of CRF features includes, but is not limited to:

- The token as is, the lowercase token, its position in the line, its shape and type according to a set of predefined classes (e.g. for shape: "UUDDDD" for "AD1900" meaning two uppercase characters and four digits. For classes, in this case we would have "AllUpperDigits", "InitUpper").
- Suffixes and Prefixes from 1 to 4 characters included.
- A set of indicator features: for example, if the token contains two digits, if four digits, if it could be an abbreviation or contain Roman numbers, etc.
- The reference style category (unique combination of class and family).
- The specific token tag, only for model 2.

A validation set containing 25% of the annotated references on which all final evaluations are based, is initially put aside and never used for training. On the remaining 75%, a set of cross-validation experiments have been performed, in order to find the best parameters and combinations of training approaches. These modifications have been tested: 1) reducing the number of features by removing the token and its lower-case version, plus all suffixes and prefixes; 2) removing references to primary sources; 3) training separate models for each of the 6 families of reference styles; 4) splitting the training data in different sizes (sets of references to parse contiguously); and 5) changing the order of the parsing tasks. Test 2 yields positive improvements and was kept, test 4 gave a window of slices of text containing 5 references as optimal for splitting annotated pages for training. Tests 1 and 5 are negative as they slightly reduced performance. Eventually, test 3 produces over-fitted models, or models that are not able to generalize properly on test data, probably due of the lack of balanced annotated data for every family. Nevertheless, removing the reference style category as a feature in the models, or using just families, leads to a slight downgrade of performance too. These details of the ablation analysis are omitted for brevity.

Once the tasks are so defined, the best configuration of CRF parameters is explored. Using a quasi-Newton gradient descent method (L-BFGS), there are two main parameters: c1 for L1 and c2 for L2 regularization respectively. Good parameters are found to be:

Table 2: Extraction results for task 1: parsing.

Class	**Precision**	**Recall**	**F1-score**	**Support**
0) null	0.679	0.553	0.609	9033
1) pagination	0.900	0.905	0.902	811
2) publisher	0.780	0.688	0.731	1029
3) author	0.847	0.862	0.855	5464
4) title	0.839	0.911	0.873	18'834
5) pub. nbr-yr	0.772	0.835	0.802	466
6) pub. place	0.860	0.873	0.867	1729
7) year	0.882	0.880	0.881	1744
avg / total	0.805	0.812	0.806	39'110

- Model 1, c1: 0.0289; c2: 0.0546.

- Model 2, c1: 1.53; c2: 0.002.

Intuitively, model 2 benefits from sparse regularization much more than model 1. The result is a set of 181'699 citations, 8632 of which are part of the golden set and 173'067 are newly parsed and extracted.

A 5-fold validation over the whole dataset gives an F1-score of 0.77 and 0.85 for task 1 and 2 respectively, while validation scores on the validation set are summarized in Tables 2 and 3, which should be read along with confusion matrices in Figure 5.

For Model 1 the main source of errors are null tokens (without tag). Several initially present tags have been removed due to them being either under-represented or too varied to be properly captured by the model. This explains the parser's difficulty in properly fitting the null tag, which ended-up being a refuge for oddities. Model 2 instead behaves consistently with the availability of data, meaning that abbreviated references are not as well captured as monographs and contributions. It is nevertheless important to note that begin tags mostly get mistaken for other begin tags, and the same for inside and end tags, all of which are minor errors.

Separate models trained to detect references to primary sources are preliminarily used to avoid mixing references to primary and secondary sources.

Disambiguation of references into citations via catalog look-up The catalog look-up attempts to match the metadata fields of the extracted ref-

Table 3: Extraction results for task 2: extraction and classification.

Class	**Precision**	**Recall**	**F1-score**	**Support**
0) out	0.936	0.958	0.947	4815
1) begin monograph	0.846	0.903	0.873	1349
2) in monograph	0.841	0.911	0.874	15’683
3) end monograph	0.862	0.894	0.878	1352
4) begin contribution	0.812	0.759	0.785	523
5) in contribution	0.892	0.802	0.845	10’930
6) end contribution	0.823	0.820	0.822	523
7) begin abbreviated	0.418	0.266	0.325	192
8) in abbreviated	0.418	0.362	0.388	1963
9) end abbreviated	0.325	0.193	0.242	192
avg / total	0.841	0.845	0.842	37’522

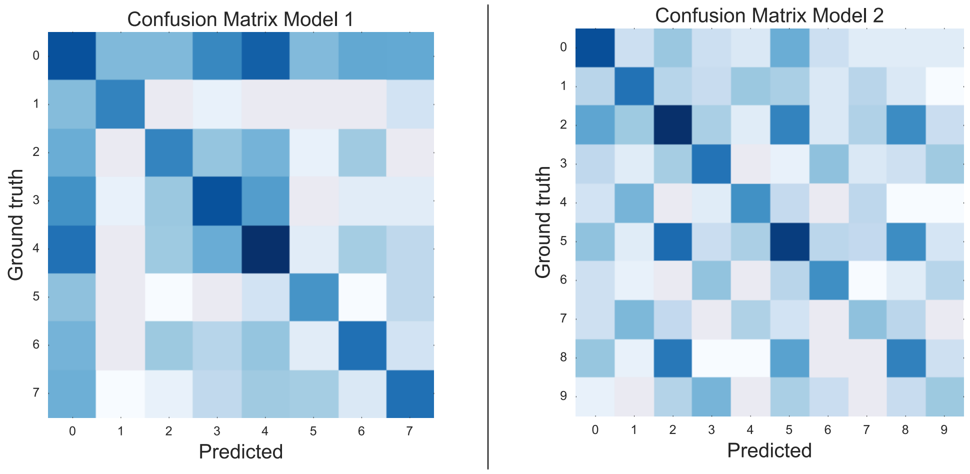


Figure 5: Confusion matrices for models 1 and 2. Identifiers are the same as in tables 2 and 3 respectively for model 1 and 2. A darker square means more matches within the bin. For example, in matrix 1, the *null* class is the most problematic as both other classes are wrongly classified as *null*, and vice-versa *null* tokens are mistaken to be of another class. In matrix 2, errors are consistent with expectations, e.g. *in monograph* mistaken for *in contribution* or *in abbreviated*.

erences against the bibliographic records contained in the SBN catalog. The look-up is performed in two steps: 1) retrieval of disambiguation candidates and 2) comparison of the input reference with each candidate.

The first step consists of retrieving a list of possible disambiguation candidates by searching through the catalog. By doing so the search space is reduced in order to avoid comparing the input reference with a high number of totally unrelated catalog records. Several experiments allowed to find the ideal settings to strike a good balance between efficiency and accuracy (and especially recall). The best results are obtained when searching for candidate records whose title contains the first two (content) words of the title in the reference, and then pruning the list, sorted by title similarity, to return a maximum of 2000 candidates. In fact, using the full cited title to search the catalog leads to an extremely low level of recall.

The second step of the look-up consists of comparing each retrieved candidate with the input reference in order to compute a global similarity score (ranging from 0 to 1): only the candidates with score above a certain threshold are then considered as matches and returned. The metadata fields that are considered for comparison are: author, title, publisher, year and place of publication. For each of these fields the similarity between candidate and input reference is calculated using fuzzy matching algorithms. Before being compared, each field is pre-processed in order to recompose hyphenated words, and remove punctuation signs as well as stop words.

Evaluation A gold standard corpus consisting of 2000 randomly sampled citations is used in order to evaluate the accuracy of the look-up: 500 manually annotated citations, and the remaining 1500 automatically extracted using the approach described in the previous section. Two annotators then disambiguate each citation by assigning the corresponding bibliographic identifier (BID) in the SBN catalog.[11]

From the original set of 2000, 4 cases are removed since the bibliographic record which corresponds to the BID assigned by the annotators is not found within our dump of the SBN catalog. 83 other citations that do not have a title are also removed, as this is the minimum requirement for a citation to be looked up. The final set of 1903 citations is used for the evaluation.

Before discussing more in detail the evaluation results, the question of what constitutes a *correct match* (for evaluation purposes) needs to be briefly

[11] The annotators use the online search interface of the library catalog to retrieve the BIDs. The search interface is accessible at `http://www.sbn.it/opacsbn/opac/iccu/free.jsp`.

Table 4: Results of the evaluation of the disambiguation module.

Class	**Precision**	**Recall**	**F1-score**	**Support**
man. annotated	0.820	0.911	0.863	493
autom. extracted	0.771	0.901	0.831	1410
total	0.784	0.904	0.840	1903

addressed as it is less trivial than it may seem. Provided that the extracted citation includes the year of publication, it is possible to try to match the citation with the record of the very same edition in the SBN catalog. As a result, a citation linked to a different edition than the referred one is considered as a wrong match. Reprints of a publication constitute the only exceptions to this rule (i.e. the BIDs of both the original version and the reprint are considered as a correct match). Moreover, in those cases where the citation points to a work in several volumes, the correct BID is the one of the record that describes the work as a whole, not the BIDs of individual volumes .

The results of the evaluation are presented in Table 4. Among the disambiguation candidates returned by the look-up, only those with a global similarity score above 0.4 are retained, as this threshold value proved to yield the best results. The candidate with the highest score is then chosen as the predicted match. For each citation, the manually assigned identifier – that may or may not be there – is compared with the identifier returned by the look-up. The accuracy of the lookup does not vary substantially between the citations that are manually annotated and those automatically extracted. While the level of recall of the look-up is overall satisfactory (0.904), its precision (0.784) could be improved.

A possible improvement of the overall accuracy concerns the matching of author names. Currently, the similarity between author names is computed by comparing the author names as they appear in a reference (mostly abbreviated) with the author field of the catalog records (where the names are always expanded). In order to boost those records that are more likely to be correct, it is possible to introduce an intermediate step where the abbreviated name form is looked up in a database of author names, and eventually replaced with the expanded form. In other words, taking into account first names for the comparison is expected to increase the capability of the look-up module of assigning a higher rank to the correct disambiguation candidate.

Finally, the evaluation results need to be interpreted in light of the following considerations. Firstly, although only the candidate with highest similarity returned by the look-up is considered, it must be noted that in a number of cases – approximately one third of the false positives – the correct match is contained in the list of candidates with similarity score above the pruning threshold (n=0.4). This detail is important given that the results of the lookup module will be eventually fed back into a correction/editing interface. Secondly, approximately 10% of the citations in our ground truth refer to publications not contained in the SBN catalog, such as publications in foreign languages or early printed books.

3.2.3 On the citation span of the corpus

This section investigates two main characteristics of the given corpus: its *cohesiveness*, defined here as the proportion of citations extracted from the corpus which refer to the corpus itself (endogenous citations), and its *connectedness*, or the dimension of the giant component in the co-citation network (considered both on the endogenous and exogenous citations, or citations to works outside of the corpus). Eventually, this section considers the feasibility of covering most of the relevant literature in a given domain using library reference monographs, and the presence of a set of core works in the spanned literature.

Cohesiveness and connectedness of the selected corpus The proposed approach rests on an hypothesis which needs testing: reference monographs in the corpus act as a hub pointing to most, or at least a considerable part of the relevant literature within the domain. Two things should happen: the selected corpus is spanning sufficiently outside of itself via direct citations, and the cited works are well-connected internally (presenting a dominant giant component in the co-citation network). If that is not the case, then reference monographs might not be effective in spanning the literature of different research areas within the same domain of study, or there might not be a significant shared literature to begin with. This hypothesis is not directly supported by (non-abundant) previous work, which in general highlights great variability in citation patterns among different disciplines in the humanities. Co-citation structures by domain and by research themes can both be found (see e.g. Ahlgren et al. [2015]; Weingart [2015]), but the proportion of citations to monographs and journal articles is quite varied in different domains [Knievel and Kellsey, 2005]. The lack of agreement over

Table 5: Citation span of the elected corpus: most citations are made to the outside.

Dataset	**Proportion**	**Matched (Extracted) refs.**
Consultation	0.0802	1861(21'337)
Not in consultation	0.0669	5398(75'270)
All	0.0699	7259(96'607)

the presence of a set of core works in the humanities would also not encourage this method (see Section 2.3). To be sure, the literature on the topic is still underdeveloped, thus the hypothesis should be tested empirically over larger datasets and across different domains.

The lookup module is initially used in order to look citations up within the corpus itself. The adapted look-up module has been manually evaluated on a small set of 500 extracted citations, resulting in a precision score of nearly 1.00 and a recall score above 0.95. High-quality results are made possible by the fact that the catalog records of the monographs in the corpus have been adapted to the task. As a reminder, the extracted citations from 700 (37%) reference monographs (of which 183, or 9.7%, are in rapid consultation) are matched against the whole corpus of 1904 (100%) books. For this purpose, only the 96'607 extracted citations in full details are considered, over the total of 181'699.

The *cohesiveness* of the corpus is defined as the proportion of extracted citations which refer to books inside of the corpus itself, over the total number of citations. Results are summarized in Table 5. Overall, only 7% of the extracted citations are made to books already within the corpus, slightly more for books on rapid consultation shelves (8%).

The *connectedness* of the corpus is equivalent to the proportion of books from the corpus which are in the giant component of the co-citation network resulting from the look-up procedure. For the dataset under consideration, the giant component is well-individuated and comprises circa 59% of the corpus. The coverage drops less than proportionally to 32.5% using only the 183 books in consultation.

These results suggest that most of the selected corpus could be useful as a source of citations pointing to the relevant literature in the domain, given that most of these citations point outside the corpus. Furthermore, a substantial part of the corpus of reference monographs is connected into the giant component of the co-citation network, suggesting the viability to span the domain.

Spanning the domain Connectedness can also be investigated at the level of the full set of extracted citations, after look-up. The resulting co-citation network comprises all the 37’626 monographs which scored sufficiently during look-up, or that are citing other monographs, and 6’030’398 edges among them. This co-citation network is not filtered by a minimum weight of edges, where the weight corresponds to the number of times two monographs have been cited together. In this network, the giant component comprises 37’359, or 99.3% of the nodes. Yet, all the monographs out of the giant component are simply part of the 700 citing monographs which are never cited, thus the giant component effectively spans all cited monographs.

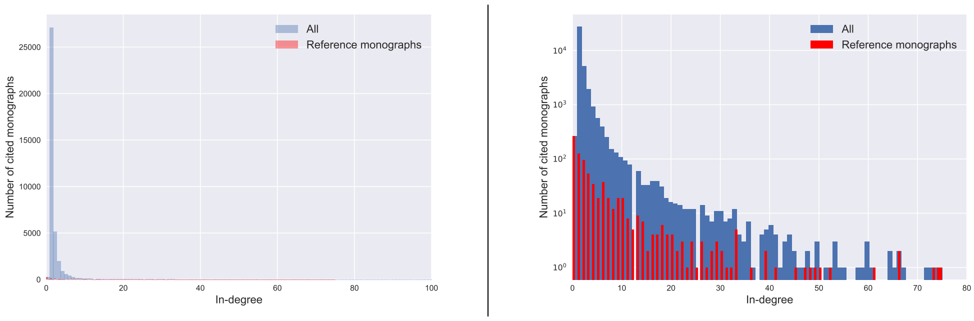


Figure 6: Distribution of in-degrees for the global extracted corpus, including reference monographs (blue/light gray) and only for reference monographs (red/dark gray). The second plot is a zoom-in to the first part of the distribution, with a log scale on the y axis.

The size of the co-citation network produced using a seed of just 700 monographs suggests at the same time that it is possible to span a wide range of works using reference monographs, and that it might be difficult to find a set of core literature. Yet, a simple look at the directed citation network connecting citing with cited books highlights a strong concentration of citations. Note that this network is not bipartite into citing and cited nodes, as a citing book (part of the 700) could also be cited in turn. This network naturally contains the same number of nodes as the co-citation network (37’626), but fewer edges (71’650, from the extracted 96’607). Every edge in this setting is a direct citation. The discrepancy from the extracted citations to the disambiguated citations is due to filtering out low confidence disambiguations and errors such as self-citations or multiple citations to the same book.

Two facts are worth noting:

1. First of all, the distribution of in-degrees is highly skewed. Figure 6 plots this distribution both globally and for reference monographs specifically, highlighting the skewness in log scale as well. In-degrees, in this settings, are the number of individual citations to a given book. 243 works (or the 0.6% of the total) are cited 20 or more times and 27'109 works (72% of the total) are cited only once.

2. The same applies to the set of the 700 citing reference monographs, which presents a division between a small group of works which are highly cited, and a larger set of barely, if ever cited works. More specifically, 37 monographs which received 20 or more citations (2% of the 1904 corpus), and providing a proportional 6.2% of given citations (4429/71'650), receive 33% of citations given to reference monographs (1280/3933). Notably, these 37 monographs are not more likely to be stored in easily accessible reference shelves (27% of them are rapid consultation, versus a proportion of 26% for the whole set of 700 reference monographs). At the same time, 1375 works in the original corpus are never cited (72% of the total).

We might conclude (and confirm) that the domain of the history of Venice is ultimately difficult to bound "bibliometrically", as the wide number of works cited from a relatively small seed suggests. Yet the field exists, as a good proportion of the cited monographs end up in the giant co-citation component. At the same time, skewed citations patterns emerge, indicating the possible existence of a core set of works in the domain, whose investigation will demand further analysis in the following chapters. Being a reference monograph indeed entails being more likely to be part of the most cited group of works, but this chance remains very low (2.5% vs 0.5%). At the same time more than 70% of the books part of the initial seed are never cited. To be sure, part of the these, and especially reference works, might be useful to the users of the library besides their relative importance in citation patterns. More problematic is the absence of most of the individuated highly cited works from the original corpus, a discovery which could inform the identification of reference works and monographs by research libraries in the future.

These results support the proposed approach, since they confirm that it is indeed possible to span a large portion of the literature in a domain to some extent and given a limited seed. In so doing, patterns of citations emerge, highlighting works that have been considered more important or durable within the community and are not, to a large extent, part of the initial corpus. No more than a few iterations of the approach sketched in Figure 2 should be necessary in order to individuate most of the core literature of a given domain, in order to integrate them in the group of references, and for the citation network to span most of the literature of interest in the domain. Existing literature is consistent in picturing the humanities as a fragmented set of disciplines, as indeed has been shown here for the historiography on Venice. At the same time, provided sufficiently focused – but certainly not big – citation data, concentrations in citation patterns do emerge, helping identifying a core set of works and spanning a relevant amount of literature in the domain. In principle, albeit necessitating further study, the proposed approach should therefore be amenable for use in all domains of the humanities.

3.3 The Scholar Index: towards a citation index for the arts and humanities

The indexation of scholarly literature is an open problem, as has been stated several times in this work already. We believe that the approach we use to index the literature on the history of Venice can be applied more generally, and be developed into a fully-fledged system. Much in the same way that national or international library catalogs are collaboratively created, every library part of this system could take responsibility for an area of scholarship of its interest. This would entail for each library to be in charge for digitization. Once done, the platform would proceed to OCR the text and mine citations, in view of their federation into a single citation index. Every library could also be responsible for the quality of the so provided citation data, by running regular evaluation and correction campaigns according to its resources. A daunting volume of work would thus be divided into more manageable chunks, and possibly distributed among several institutions. This collaborative approach to citation indexing in the arts and humanities, which we named *Scholar Index*, is illustrated in Figure 7.

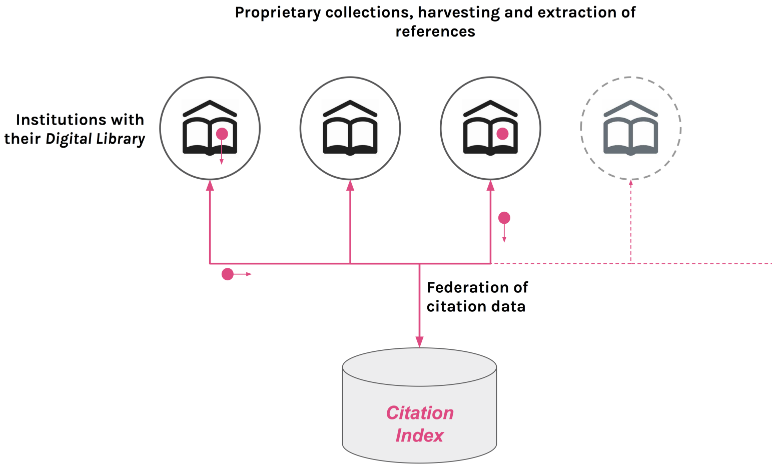
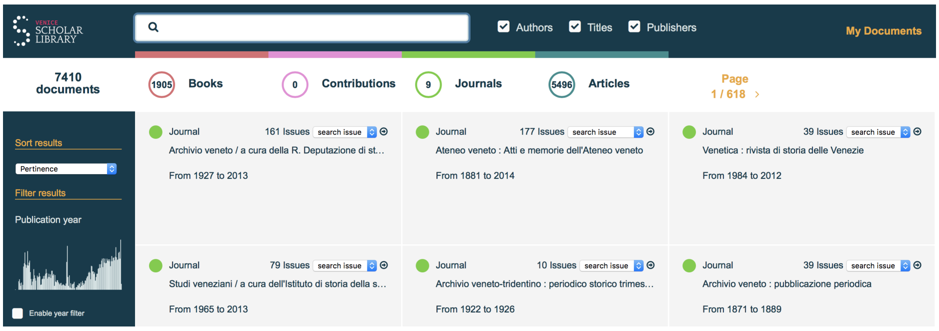

Proprietary collections, harvesting and extraction of references
Institutions with their *Digital Library*
Federation of citation data
Citation Index

Figure 7: The Scholar Index platform: a federated and collaborative citation index build by the shared effort of individual libraries, where each library feeds citation data extracted from client digital library applications.

In order to illustrate the approach, we developed two interfaces and used them with the data from Venice: a digital library and a citation index.[12] The digital library and the citation index are connected via citations. The digital library provides access to the digitized materials, and points to the index through disambiguated references, as illustrated in Figures 8, 9, 10.

[12] See Chapter 9 for details on how to access.



Journal 161 Issues search issue
Archivio veneto / a cura della R. Deputazione di st...
From 1927 to 2013

Journal 177 Issues search issue
Ateneo veneto : Atti e memorie dell'Ateneo veneto
From 1881 to 2014

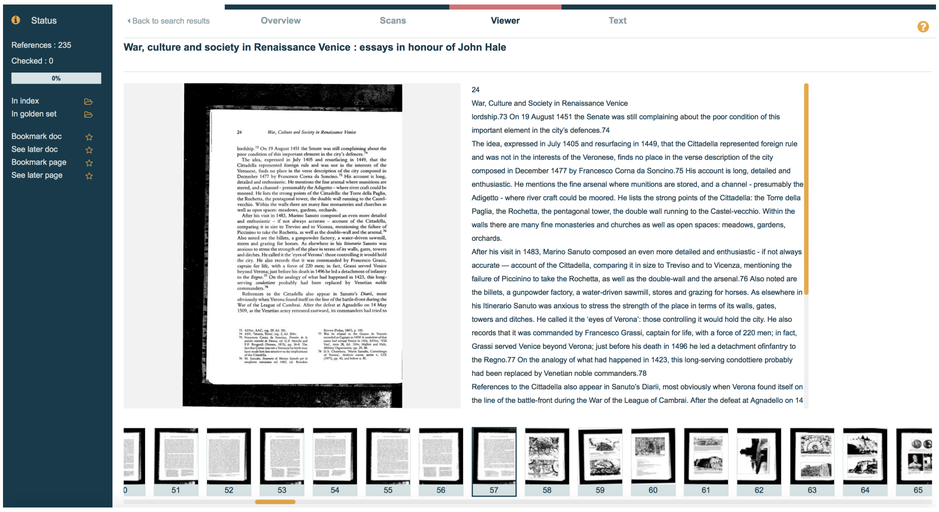
Journal 39 Issues search issue
Venetica : rivista di storia delle Venezie
From 1984 to 2012

Journal 79 Issues search issue
Studi veneziani / a cura dell'Istituto di storia della s...
From 1965 to 2013

Journal 10 Issues search issue
Archivio veneto-tridentino : periodico storico trimes...
From 1922 to 1926

Journal 39 Issues search issue
Archivio veneto : pubblicazione periodica
From 1871 to 1889

Figure 8: Search over the metadata of the digitized collection.



War, culture and society in Renaissance Venice : essays in honour of John Hale

24

War, Culture and Society in Renaissance Venice

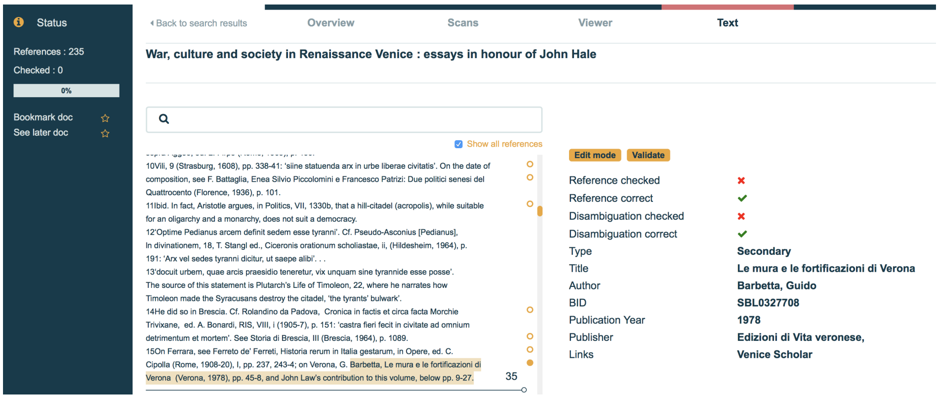
lordship.73 On 19 August 1451 the Senate was still complaining about the poor condition of this important element in the city's defences.74

The idea, expressed in July 1405 and resurfacing in 1449, that the Cittadella represented foreign rule and was not in the interests of the Veronese, finds no place in the verse description of the city composed in December 1477 by Francesco Corna da Soncino.75 His account is long, detailed and enthusiastic. He mentions the fine arsenal where munitions are stored, and a channel - presumably the Adigetto - where river craft could be moored. He lists the strong points of the Cittadella: the Torre della Paglia, the Rochetta, the pentagonal tower, the double wall running to the Castel-vecchio. Within the walls there are many fine monasteries and churches as well as open spaces: meadows, gardens, orchards.

After his visit in 1483, Marino Sanuto composed an even more detailed and enthusiastic - if not always accurate — account of the Cittadella, comparing it in size to Treviso and to Vicenza, mentioning the failure of Piccinino to take the Rochetta, as well as the double-wall and the arsenal.76 Also noted are the billets, a gunpowder factory, a water-driven sawmill, stores and grazing for horses. As elsewhere in his Itinerario Sanuto was anxious to stress the strength of the place in terms of its walls, gates, towers and ditches. He called it the 'eyes of Verona': those controlling it would hold the city. He also records that it was commanded by Francesco Grassi, captain for life, with a force of 220 men; in fact, Grassi served Venice beyond Verona; just before his death in 1496 he led a detachment ofinfantry to the Regno.77 On the analogy of what had happened in 1423, this long-serving condottiere probably had been replaced by Venetian noble commanders.78

References to the Cittadella also appear in Sanuto's Diarii, most obviously when Verona found itself on the line of the battle-front during the War of the League of Cambrai. After the defeat at Agnadello on 14

Figure 9: Viewer: allows to read a publication with image and text side by side. This is particularly important in order to appreciate the quality of the OCR.



References : 235
Checked : 0
0%

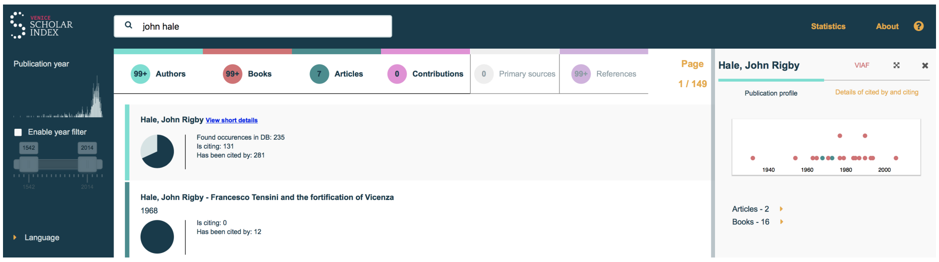
War, culture and society in Renaissance Venice : essays in honour of John Hale

10Vili, 9 (Strasburg, 1608), pp. 338-41: 'siine statuenda arx in urbe liberae civitatis'. On the date of composition, see F. Battaglia, Enea Silvio Piccolomini e Francesco Patrizi: Due politici senesi del Quattrocento (Florence, 1936), p. 101.
11Ibid. In fact, Aristotle argues, in Politics, VII, 1330b, that a hill-citadel (acropolis), while suitable for an oligarchy and a monarchy, does not suit a democracy.
12'Optime Pedianus arcem definit sedem esse tyranni'. Cf. Pseudo-Asconius [Pedianus], In divinationem, 18, T. Stangl ed., Ciceronis orationum scholiastae, ii, (Hildesheim, 1964), p. 191: 'Arx vel sedes tyranni dicitur, ut saepe alibi'. . .
13'docuit urbem, quae arcis praesidio teneretur, vix unquam sine tyrannide esse posse'. The source of this statement is Plutarch's Life of Timoleon, 22, where he narrates how Timoleon made the Syracusans destroy the citadel, 'the tyrants' bulwark'.
14He did so in Brescia. Cf. Rolandino da Padova, Cronica in factis et circa facta Morchie Trivixane, ed. A. Bonardi, RIS, VIII, i (1905-7), p. 151: 'castra fieri fecit in civitate ad omnium detrimentum et mortem'. See Storia di Brescia, III (Brescia, 1964), p. 1089.
15On Ferrara, see Ferreto de' Ferreti, Historia rerum in Italia gestarum, in Opere, ed. C. Cipolla (Rome, 1908-20), I, pp. 237, 243-4; on Verona, G. Barbetta, Le mura e le fortificazioni di Verona (Verona, 1978), pp. 45-8, and John Law's contribution to this volume, below pp. 9-27.
35

Reference checked	✗
Reference correct	✓
Disambiguation checked	✗
Disambiguation correct	✓
Type	Secondary
Title	Le mura e le fortificazioni di Verona
Author	Barbetta, Guido
BID	SBL0327708
Publication Year	1978
Publisher	Edizioni di Vita veronese,
Links	Venice Scholar

Figure 10: Text view: allows to full-text search within a publication, highlighting all extracted references and links to the relative entries in the index.

The citation index provides instead no access to full-contents, also for reasons of copyright, but allows for the exploration of the network of citations, as shown in Figures 11, 12, 13.



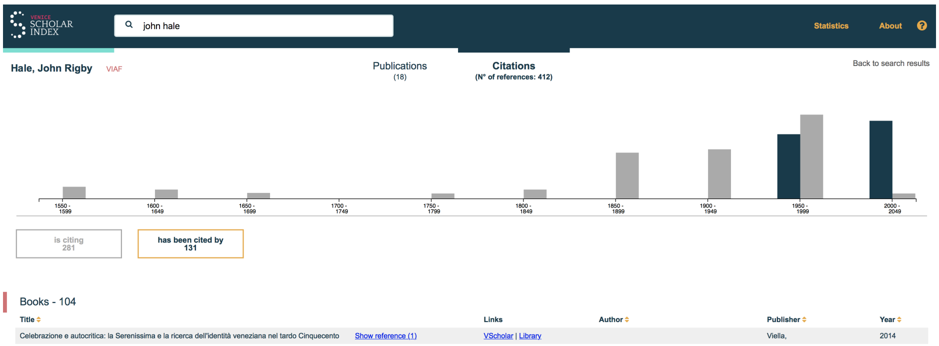
john hale

Hale, John Rigby View short details
Found occurences in DB: 235
Is citing: 131
Has been cited by: 281

Hale, John Rigby - Francesco Tensini and the fortification of Vicenza
1968
Is citing: 0
Has been cited by: 12

Hale, John Rigby

Figure 11: Search results: citation data is aggregated by author, publication or primary source, with full-text access to the text of extracted references. Search results are conveyed along with their relevant citation information (citations made and received, publications for an author). Authors are linked to the Virtual International Authority File (VIAF) repository.



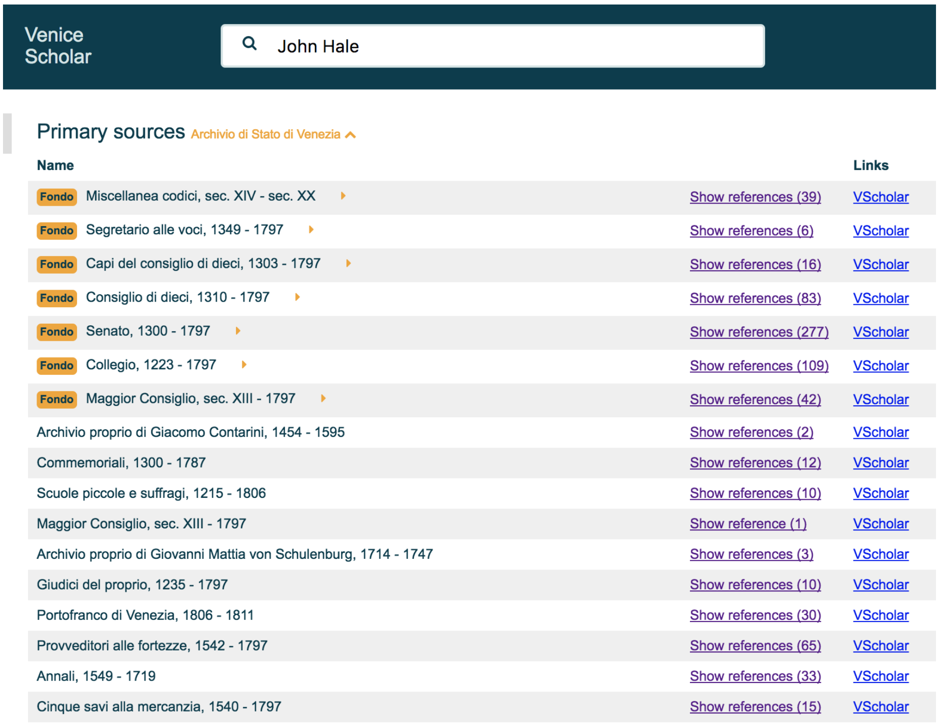
john hale

Hale, John Rigby

Books - 104

Title ⬍		Links	Author ⬍	Publisher ⬍	Year ⬍
Celebrazione e autocritica: la Serenissima e la ricerca dell'identità veneziana nel tardo Cinquecento	Show reference (1)	VScholar | Library		Viella,	2014

Figure 12: Citation timeline: every aggregated entity has a dedicated page with a timeline of citations (made and received), and a list of relevant sources.



Venice Scholar

John Hale

Primary sources Archivio di Stato di Venezia

Name		Links
Fondo Miscellanea codici, sec. XIV - sec. XX	Show references (39)	VScholar
Fondo Segretario alle voci, 1349 - 1797	Show references (6)	VScholar
Fondo Capi del consiglio di dieci, 1303 - 1797	Show references (16)	VScholar
Fondo Consiglio di dieci, 1310 - 1797	Show references (83)	VScholar
Fondo Senato, 1300 - 1797	Show references (277)	VScholar
Fondo Collegio, 1223 - 1797	Show references (109)	VScholar
Fondo Maggior Consiglio, sec. XIII - 1797	Show references (42)	VScholar
Archivio proprio di Giacomo Contarini, 1454 - 1595	Show references (2)	VScholar
Commemoriali, 1300 - 1787	Show references (12)	VScholar
Scuole piccole e suffragi, 1215 - 1806	Show references (10)	VScholar
Maggior Consiglio, sec. XIII - 1797	Show reference (1)	VScholar
Archivio proprio di Giovanni Mattia von Schulenburg, 1714 - 1747	Show references (3)	VScholar
Giudici del proprio, 1235 - 1797	Show references (10)	VScholar
Portofranco di Venezia, 1806 - 1811	Show references (30)	VScholar
Provveditori alle fortezze, 1542 - 1797	Show references (65)	VScholar
Annali, 1549 - 1719	Show references (33)	VScholar
Cinque savi alla mercanzia, 1540 - 1797	Show references (15)	VScholar

Figure 13: Citations to primary sources: the index also links to external collections of primary sources, in this case documentation at the Archive of Venice. Citations to any level of the archival hierarchy are provided, following its structure. The user can easily move from a publication to a document series and see all publications which referred to it, over time.

The Scholar Index platform is thus able to aggregate citation data from many library collections into a unique system, allowing users to not only navigate the resulting network, but also have improved access to collections of primary sources such as archives. This platform is currently being tested with historians, archivists and librarians, and will be soon released for the community. Another work in progress is its connection with Open Citations,[13] in order to ingest all our citations into that repository for dissemination.

[13] A triple store of openly available citations, accessible at `http://opencitations.net`.

3.4 Discussion

In this chapter we suggested that, in order to map the intellectual organization of history as a discipline, novel datasets are needed, offering unprecedented coverage in terms of publication and source typologies and over time. We proposed to look for a solution in research libraries, where coherent collections of books are curated for the specific needs of scholars. We focused on a case-study, the historiography on Venice, and developed an approach to produce a representative citation dataset from scratch. This approach relies on the use of reference works and monographs in a specialized research library, which are used as a seed to extract citations and, through them, span the relevant literature of the field. Testing our approach required the development of a technical pipeline, of which we gave full details.

This work allowed to produce new citation data, which will be used in what follows, and to confirm that library collections can indeed provide for a focused way to generate such citation datasets in the arts and humanities. Furthermore, we proposed a generalization of our approach named the *Scholar Index*: a platform to create a collaborative and distributed citation index in the arts and humanities, harnessing the collective work of research libraries. A working prototype was developed for the Venetian case-study, as an illustration.

It is worth noting that, from our experience, several challenges remain open when facing the task of building citation indexes in the arts and humanities. More specifically we have:

1. *Intrinsic issues*: the fragmentation and variety of the literature, in terms of languages, venues of publication, cited sources; the lack of incentives and the parallel raise of alternative monitoring approaches such as altmetrics; the lack of central actors either private or public which see a social or economic incentive to index this literature.

2. *Infrastructural issues*: the poor digital availability of the literature; the difficult and at times impossible access to catalog metadata (of libraries, archives, museums, etc.), which hinders the interconnection of collections via citations; the still tentative adoption of standards to expose metadata uniformly, and the limited availability of APIs for their harvesting.

3. *Legal issues* for the use and publication of citation and publication data. Issues related to copyright on digitized materials constitute a

great obstacle to citation mining, and text mining of scientific publications in general. In particular, legal frameworks might limit the freedom to openly share citation data, as well as datasets for training and evaluation of citation mining solutions. This issue applies not only to historical materials, but especially to contemporary scholarly publications.

4 Mapping the intellectual organization of historians

The main research questions raised at the beginning of this work include: how the humanities are organized intellectually? How the knowledge they produce accumulates? How the increasing volume of publications and the advent of the Digital Humanities are affecting the ways scholars in the humanities conduct and publish their research? All these questions are directly related to the challenges of scholarly information retrieval, on one side, and research evaluation, on the other side. History is, in this respect, a particularly compelling example. Often seen as a boundary discipline in-between the social sciences and the humanities, history is characterized by deeply-rooted intellectual traditions and a practically open-ended wealth of primary sources to rely upon, determining a strong grounding in space and time.

This chapter takes an exploratory approach through the methods of science mapping (see Section 2.5), considering a specific bibliometric level of analysis: citations between books. As discussed in Section 2.3, books are still the most important publication typology in the humanities, a fact that shows no sign of change in recent years. At the same time, their "citation profile" is poorly understood given that only recently citation indexes such as the Web of Science and Scopus have started indexing them. We consider here a recent and representative set of books on the history of Venice, transformed in a citation dataset as discussed in Chapter 3, in order to map the intellectual landscape and current trends in this research area. We further explore the most cited books, that is to say the core literature, in order to qualify it and discuss its structural role, with the goal of uncovering how historians relate to previous literature. The purpose of this chapter is also that of setting the ground by rising the key questions that will be explored in the rest of this work.

It is worth noting that the history of Venice offers a representative example of a research area in historiography. Relying on two hundred years of erudite scholarship, just to consider modern times [Dursteler, 2013], and often mixed with political or ideological motivations [Infelise, 2002; Povolo, 2002], the most recent historiography on Venice is inevitably conditioned by its past. At the same time, and like many other research areas in history and beyond, Venice saw a surge in internationalization during the past few decades, effectively managing to connect its local community to other, mostly English and French-speaking ones [Grubb, 1986; Davidson, 1997]. As a consequence, studies proliferate and new avenues of research are being opened with increased frequency. Venice can effectively be considered as a playground, representative of the most recent trends in historiographical research [Horodowich, 2004]. In this context, it appears not at all trivial to ask the question on how the intellectual landscape of the historians of Venice is organized, given the novelty of recent scholarship, but also its need to dialog with the past in order to forge its identity.

We will use the terms domain, field or area of research interchangeably in what follows, unless otherwise specified. We will also refer to books generally to indicate a variety of publications that take this form, including scholarly monographs, edited volumes, edited primary sources, reference works, among others. This chapter borrows substantially from Colavizza [2017a].

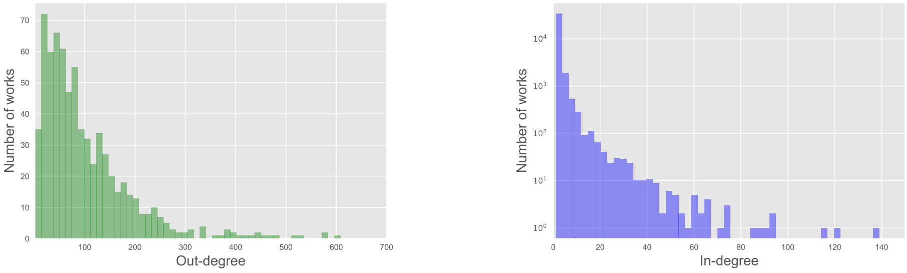
4.1 Methods and data

There are perhaps two main challenges when considering intellectual landscapes in the humanities, as mapped by citations: individuating a representative sample of the literature of a given community, and acquiring its citation dataset. In the absence of comprehensive book citation indexes [Zuccala et al., 2015], the only viable option is to use the resources available from research libraries and the advice of domain experts in order to delineate a first sample of works, extract their citations and then proceed to enlarge the corpus iteratively. We consider here a published citation dataset from books to books, on the history of Venice [Romanello and Colavizza, 2017][14]. The citing set of books was selected trying to cover on-demand works, aiming at representing recent trends in the field, including the tightly connected areas of the histories of art and architecture. Different means were used in order to individuate these books, among which the shelving strategy of the library (selecting works in rapid consultation shelves), catalog classification and scholarly bibliographies. Furthermore, only books with reference lists were considered in order to extract their references, thus there is no ambition of comprehensiveness. To be sure, this selection did not entail specific biases by publisher or date of publication. As a consequence, the dataset only considers book to book citations, irrespective of the frequency of in-text references, therefore resulting in an unweighted directed citation network. The

[14] More details are provided in Chapter 3, Section 3.2 and in Colavizza et al. [2017].

exclusion of journal articles is partially justified by the fact that they likely do not become part of the core literature (see Hammarfelt [2011]; Colavizza [2017b] and Chapter 5 below).

The dataset comprises 700 citing books and 37'362 cited books. 264 citing books are never cited in turn. The total number of individual citations (citing to cited) is 73'268, or slightly more than 100 for every citing book. The distribution of the number of citations made by these 700 books is given in Figure 14a for reference. Values are reasonably between 20-30 and 300, with some more extensive but rare reference lists. The distribution of the received citations is, instead, more skewed, as shown in Figure 14b. In particular, 27'109 works are cited only once, and just 769 ten or more times. We consider this last group of books to be the core literature which will be discussed in what follows.

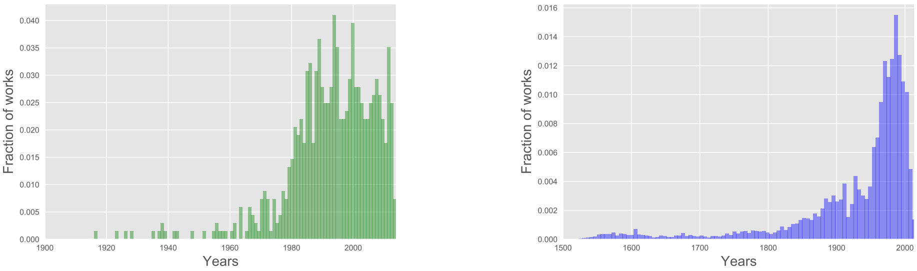


(a) Number of given citations from citing works (out degree).

(b) Number of received citations by cited works (in degree). The y axis is on log scale.

Figure 14: The distribution of the given and received citations (or the out and in degrees of the directed citation network). The distribution of the number of received citations is particularly skewed. Only 769 works are cited ten or more times, and constitute the core literature. Please note the scales in two two plots differ significantly.

The age of cited books is given in Figure 15b. The age of some cited works is very considerable, with publications dating back to the Renaissance. Some turning points in the historiography on Venice also emerge, notably the end of the Republic of Venice in 1797 and two world wars, which determined a reduction in the number of new publications, in the latter case common to all domains of science [De Solla Price, 1965]. Besides, the volume of cited literature rises considerably moving closer in time, another phenomenon in common with the sciences. The distribution of the age of citing books is, instead, concentrated for the most part between the years 1980 to 2013, as shown in Figure 15a. This effect results from the selection choice which was made, namely using library resources to recover the most relevant and up-to-date reference publications on the history of Venice. The citing group is thus representing recent historiography, and is relatively up to date at least by humanities' standards, as intended.



(a) Age of citing works.

(b) Age of cited works.

Figure 15: The distribution of the age of the citing and cited works, respectively. Citing works mainly concentrate from 1980 to 2013, while cited works essentially span from the Renaissance to the present day.

The languages and places of publication of the citing and cited works are given in Table 6. Italian is by and large the most represented language, followed by the main Western languages. The dataset thus strongly represents local as well as international historiography on the topic, and confirms the tendency of scholarship to rely heavily on research published in national languages.

Networks commonly follow the terminology of graph theory, and are thus made of nodes (or vertices) connected by edges. In our case, nodes are books and edges are citation relations among them. Edges can also be weighted in order to distinguish between stronger and weaker relations. In this chapter, three kinds of citations networks will be used, which are all often used in order to map different aspects of the intellectual structure of a research area or discipline.[15] The most basic one is the directed, not weighted citation network

[15] With respect to the full dataset, 27 citing books have been removed as duplicate

Table 6: Place of publication and language for most of the cited and citing sources. This information comes from library catalog metadata.

	Pb. country (citing)		**Pb. language (citing)**		**Pb. country (cited)**		**Pb. language (cited)**
IT	540	**ita**	520	**IT**	24'151	**ita**	23'052
GB	46	**eng**	112	**GB**	3256	**eng**	6407
US	45	**fre**	37	**FR**	3241	**fre**	3782
FR	25	**ger**	25	**US**	2635	**ger**	2203
DE	24	**lat**	4	**DE**	2055	**lat**	1256

where every node is a book, and an edge exists from one node to another if the former cites the latter. This network comprises 37'200 nodes and 68'748 edges. Given this representation, two other networks can be constructed. The bibliographic coupling network is a weighted, undirected network where every node is a citing book, and every edge represents the overlap of references between the two books [Kessler, 1963]. For example, if two books both refer to the same three books, they will be connected by an edge of weight 3. This network comprises 673 nodes and 87'419 edges, and accounts for how recent literature defines an intellectual landscape according to its use of the literature. The co-citation network is a weighted, undirected network where every node is a cited book, and an edge is established between two nodes if the two are cited together in the same reference list [Marshakova Shaikevich, 1973; Small, 1973]. The weight of the edge is given by the number of times the two books were cited together in different reference lists. A minimum weight of 2 is established here as a threshold, in order to filter-out books cited only once or anyway weak and possibly episodic relations. This last network comprises 9061 nodes and 288'782 edges among them, and accounts for the way the literature of the area was used by recent scholarship. Recent trends in the literature can be mapped by bibliographic coupling, the literature or 'intellectual base' they rely on by using co-citation networks [Persson, 1994; Hammarfelt, 2011].

editions, despite the fact that most of these editions constitute updates from a previous work, due to the fact that even a revised or extended edition is likely to contain substantial overlaps with previous ones in terms of references. When multiple editions of a work exist, the most recent one is kept; when translations of a work exist, the original is kept, but if the translation also includes an updated version of the work, it is retained instead.

4.2 The recent historiography on Venice

Our starting point is the topology of the bibliographic coupling network. The books part of the sample were selected considering a broad definition of historiography, also including the histories of arts and architecture. It is therefore important to assess to what extent the citation network at the book level allows to characterize the field as a whole (i.e. the network is connected or not) and individuate its sub-areas and topics of interest through clustering (i.e. community detection).

In order to qualify the results of any clustering, all nodes (citing books) have been classified with a unique keyword corresponding to their general sub-area (history, arts or architecture), and with two groups of keywords (every book can receive no, one or more keywords for these two groups, as appropriate) for topics and periods under consideration. This classification has been performed manually by experts. It relies on the Dewey and subject classifications of the Italian National catalog, which could not be directly used due to its granularity being either too generic or too specific in the dataset at hand. It should be noted that manual classification of publications, in itself important to interpret results, is perhaps the least scalable part of the whole study. The resulting classification is made available for inspection (see Data Availability). There are 419 books classed under history, 129 arts and 125 architecture. 42 keywords for topics include for example "social" history (86 books), "politics" (80), "individuals" (62) and "churches" and religion institutions (52), "urban" life and architecture (45). 29 books could not be classed with topic keywords. The keywords for the periods under consideration are the Renaissance (234), eighteenth century (161), seventeenth century (149), the middle ages and late ancient period (122), nineteenth century (85) and more recent times (20). 190 books could not be clearly classified by period. It is clear at glance that the historiography on Venice has a strong focus on the early modern period, especially the Renaissance, with less attention given to the periods of the early middle ages – likely due to the lack of sources – and the modern period – likely in part due to the over-abundance of sources –, and covers a great variety of topics, both established since a long time or emerged recently. This classification according to the library catalog can both provide a direct clustering of books into communities, and serve as a way to qualify – but not evaluate – the results of automated clustering using citation data. In particular, frequent keywords and periods can help qualify clusters much in the same way the most significant words in a topic model can help assign a label to it. It must be stressed that the two perspectives, library classification and clustering based on citations, need not coincide.

One of the most important perspectives of analysis on networks are communities, or clusters of nodes. Despite the absence of an agreed upon definition of network communities, we can say intuitively that they are clusters of nodes which are more likely to be connected among themselves than with the exterior [Fortunato and Hric, 2016]. Several methods exist for the detection of communities in networks, and their application to citation networks has been extensively explored [Šubelj et al., 2016]. One particularly popular method relies on modularity maximization [Newman and Girvan, 2004], and has eventually been extended to incorporate a resolution parameter, helping to tune the size and thus resulting number of clusters [Reichardt and Bornholdt, 2006]. Fast implementations exist, among them the Louvain algorithm [Blondel et al., 2008] is a well known one. This method has its features – for example it is not deterministic, thus different runs can yield different results – and shortcomings [Fortunato and Barthelemy, 2007; Good et al., 2010], it is therefore important to compare it with other methods or use external information to interpret any clustering result. Yet, modularity maximization gave by far the most interpretable results on the dataset under analysis here, where several other methods even failed to distinguish any structure in the network, mainly due to its density. The interested reader can experiment using publicly available code.[16] In absence of further specification, when a clustering solution is discussed it is one of the possible similar results from modularity maximization.

At large scale, and despite considering quite different sub-areas such as arts and history, the network is almost connected – i.e. only two nodes are not part of its giant component. The giant component is the largest set of connected nodes (a set of nodes is connected if a path exists among any two of them), and in this context can be interpreted as the largest group of publications with some reference in common, possibly indirectly. The network is also very dense: it contains almost 40% af all possible edges among nodes, entailing that a strong overlap exists across the reference lists of historians. Such well-connected network inevitably brings some difficulty in finding clusters of nodes. A comparison of a clustering solution with the

[16] Most analyses relied on igraph [0.7.1] [Csardi and Nepusz, 2006] and Vincent Traag's community detection library [0.5.3] available at `https://github.com/vtraag/louvain-igraph`. The code to replicate all analyses and plots of this paper is available online, see Chapter 9.

general categories from the library catalog is given in Figure 16. The labels of the clusters in Figure 16b have been given inspecting the general categories, keywords and periods of the books within each cluster. With respect to this dataset, the field appears intellectually organized according to two main sub-areas, namely history on one side, arts and architecture on another side, plus over the dimension of time, according to the main periods of interest for the historians of Venice. Most notably, the history of the early modern Republic, especially its Renaissance period, is the focal point of attention by number of publications. To a lesser degree the middle ages, and to a much lesser degree the nineteenth century and beyond. Borderline smaller areas of activity, such as the applied arts, emerge as well at this level.

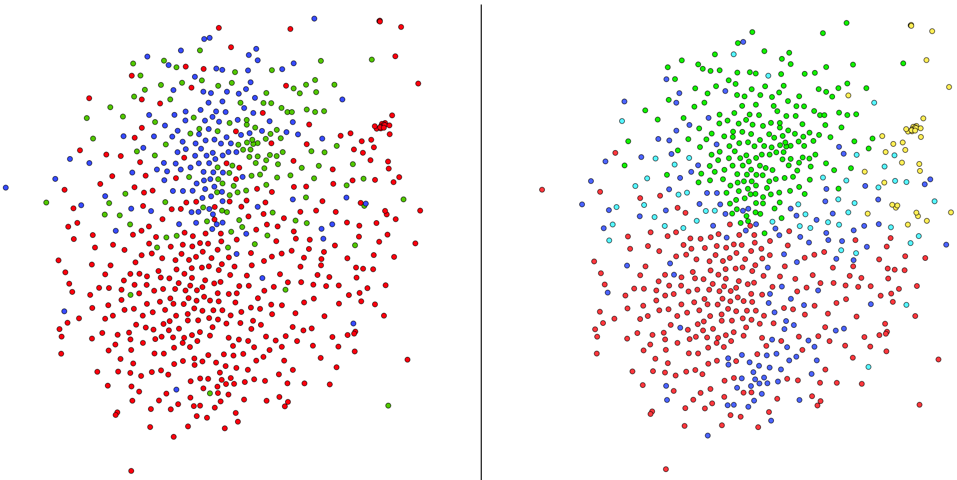
Starting from this most general situation, finer-grained clusters can be determined, either by tuning the resolution parameter or by further clustering an already individuated cluster. By further clustering the largest history cluster in Figure 16b (in red), a set of smaller clusters emerge, which we might consider as broad areas of interest of the recent literature. Four clusters relate to the Renaissance period, from different perspectives:

1. Aspects related to the political history and the elites, touching on foreign relations and the Venetian empire. An example is Donald Queller's "Il patriziato veneziano: la realtà contro il mito" (The patriciate of Venice: reality vs myth).

2. Social and religious history, also touching upon censorship, gender and culture. Examples are Satya Datta's "Women and men in early modern Venice: reassessing history" and Muir's "Civic ritual in Renaissance Venice". Most of the publications in this cluster are quite recent, after the year 2002.

3. The government of the city and its Mainland state. For example, Claudio Povolo's "L'intrigo dell'onore: poteri e istituzioni nella Repubblica di Venezia tra Cinque e Seicento" (The intrigues of honor: powers and institutions in the Republic of Venice between the sixteenth and seventeenth century).

4. Economic history. E.g. Richard Rapp's "Industry and economic decline in seventeenth-century Venice". This is a quite old cluster dating back mostly to the 1970s and 80s.



(a) The clusters according to catalog metadata at the least granular level available. Red/grey: history, blue/darker grey: history of architecture, green/lighter grey: history of the arts.

(b) The clusters from to modularity maximization. Red: early modern history, green: arts and architecture, blue: history of the middle ages, cyan: nineteenth century history, yellow: applied arts.

Figure 16: Different clustering of citing books (giant component of the bibliographic coupling network) according to catalog metadata (left) and citation information (right). At this level, citation information captures general categories and the periods under consideration, as well as smaller sub-areas such as applied arts (yellow on the right). This visualization was made with Gephi 0.9.1 [Bastian et al., 2009], using Force Atlas 2 with default parameters but for LinLog mode, scaling 0.5 and edge influence 0.8. Edges are omitted: the network is connected. It is important to note that the disposition of the nodes is related to, but is not determined only by clusters found by maximizing modularity [Jacomy et al., 2014].

Another cluster is instead made by works devoted to the eighteenth century, with a mix of perspectives spanning from politics and the role of elites, to the reform of government or the social and cultural aspects of the period. An example is given by Volker Hunecke's "Il patriziato veneziano alla fine della Repubblica: 1646-1797" (The Venetian patriciate at the end of the Republic). Besides exceptions, all clusters include relatively recent works (the 1990s and 2000s for the best part).

The bulk of the historiography on early modern Venice is a mix of old topics often reconsidered under new perspectives. Most of these areas of interest of the recent literature have a long tradition of study among historians of the city [Grubb, 1986; Davidson, 1997; Dursteler, 2013]. In fact, their emergence in the network signifies the presence of a continuity in the use of the literature. Remarkable is instead the relative lack of recent efforts in the study of the economic history of Venice, at least within the dataset. The social and economic history of Venice had its heydays during the 1960s and 70s, mainly due to the influence of the works of Fernand Braudel and the *École des Annales*, but it has become since then of less importance. The most notable novelty in this recent historiography – as already discussed in the literature [Horodowich, 2004] – is the surge in the number of important studies dealing with a new social history, marking 'a shift in interest from order to disorder, from orthodoxy to dissent, from the centre of power to the broader social context' [Davidson, 1997]. Examples in this respect are the relatively new trends of gender and women history.

With respect to the histories of arts and architecture cluster, the division is simpler and historically more stable: architecture broadly organizes itself into a urban dimension, where palaces and the city more in general are considered, and a dimension related to religious buildings, especially churches and convents. The arts are instead largely dominated by the study of individual painters and their schools, with a division by period into the Renaissance and the later eighteenth century and beyond. Interestingly, in this later period, more attention is given to private collecting, whilst in the previous period applied arts such as jewelry have an influence due to the proximity with the middle ages, when painting played a subordinate role.

The middle ages cluster generally orbits around two dimensions too: the history of the establishment of the Venetian empire, with strong focus on its commercial as well as political dimensions, and the history of the urban development of the city and its relation to the lagoon and its natural environment. The works of John Julius Norwich ("Venice: the rise to empire") and Gerhard Rösch ("Venezia e l'Impero", Venice and the Empire) feature among the former group; Wladimiro Dorigo's "Venezia romanica: la formazione della città medioevale fino all'età gotica" (Romanesque Venice: the formation of the medieval city until the Gothic period) is the most important work in the latter, and one that could have fitted into the architecture cluster as well.

Lastly, the nineteenth century cluster is recently mainly devoted to the social, cultural and political history of the city after the fall of the Republic.

Another older cluster deals with the events of the year 1848, when a short-lived Republic was established between two periods of Austrian domination. All these considerations evidently apply only with respect to the sample under consideration.

Despite the fact that a relatively clear landscape of the recent historiography on Venice emerges from citation network at the level of books, it must be noted that the quality of the clustering, as measured by the modularity of the partitions, as well as by direct inspection, rapidly degrades while rising the number of clusters. The bibliographic coupling network among citing books is very well connected, and effectively provides a broad overview of the field, without allowing for a too fine-grained individuation of small clusters, whose emergence might require further information, such as citations to journal articles and primary sources. This specific citation landscape relies for its tight organization on the use of previous literature, or the intellectual base of the historians of Venice. It is possible to consider the previously introduced co-citation network, in order to explore how such literature has been used, and relate it to specific clusters of citing books.

4.3 The intellectual base and its core

The co-citation network, filtered to include only edges with weight of two or more, is again an almost connected graph (only 73 our of 9061 nodes are not part of the giant component). Furthermore, its density is much lower than for the bibliographic coupling network, at 0.007. This follows directly from the fact that most of the literature is cited but a few times. In order to highlight the role of the core literature in the co-citation network, three centrality measures at the node level are considered:[17]

- *Betweenness*: accounts for the capacity of a node to bridge different areas of the network, which would be less well-connected without it.

- *Local clustering coefficient*: the proportion of neighbor nodes which are connected in turn. A neighbor node is directly connected to the node of interest. If all the neighbors of a node are connected among them, its local clustering will be 1. It gives an idea on how densely connected the local neighborhood of a node is, and allows to probe for the presence

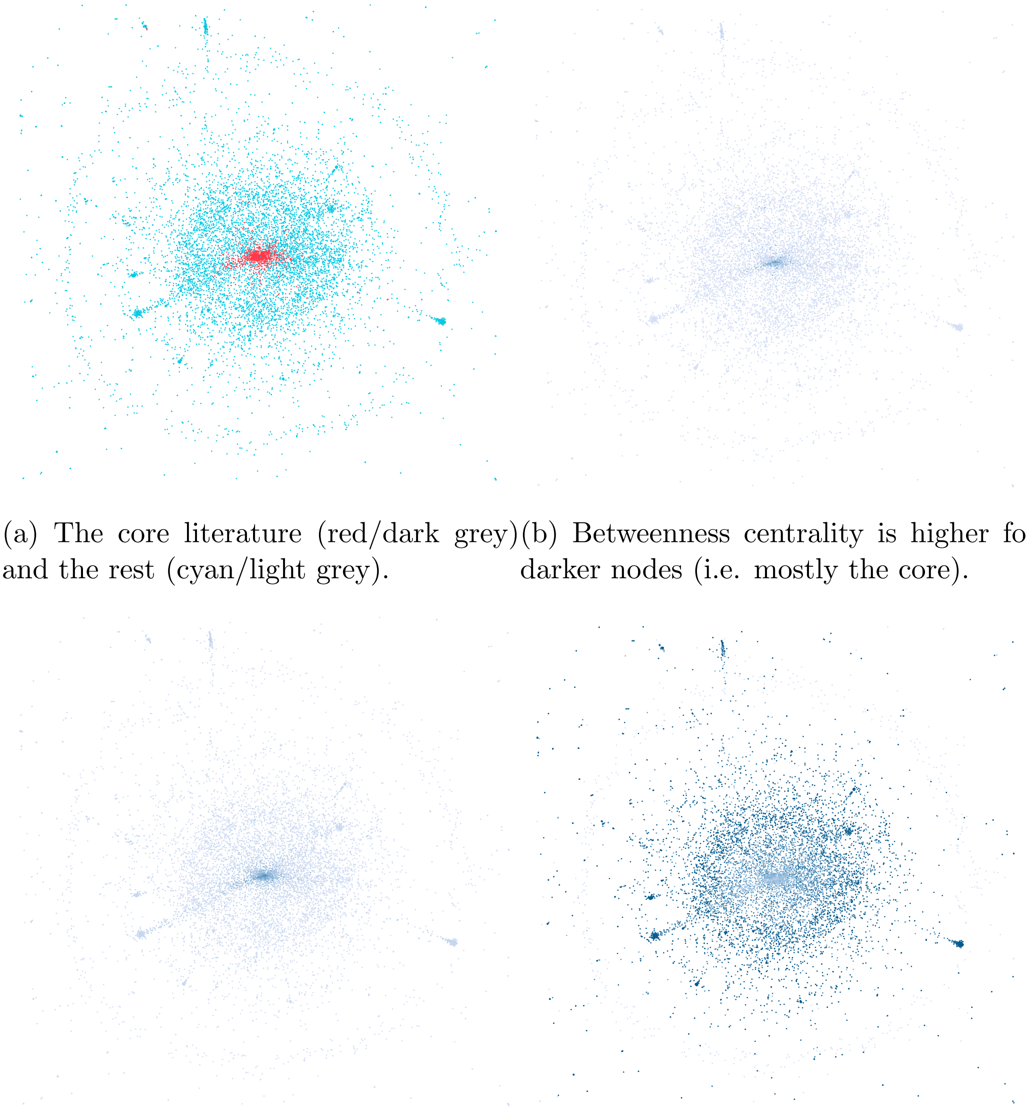
[17] For formal definitions see Newman [2010].

of structural holes, or areas of a network with missing links, giving an important role to local brokers [Newman, 2010, 202].

- *PageRank*: accounts for the importance of a node with respect to it being connected to other important nodes. PageRank is not a centrality measuring the intermediation capacity of nodes, by their global role as mapped by the recursive relative importance of their neighbors. It is used here for comparison with the other two centralities.

These three measures play together: it is expected that betweenness and PageRank will be high, and local clustering will be low for the core literature. Intuitively, this would mean that the core literature is able to connect different areas of the network, thus groups of works that have been cited by different communities (this entails high betweenness and low local clustering) and is particularly connected among itself (high PageRank) due to the fact that core works are frequently cited together.

Figure 17a displays the giant component of the co-citation network, and highlights the core literature into it for reference (in red/dark gray). With this picture in mind, it is possible to appreciate how the intuitive role of the core finds confirmation using the three proposed measures of centrality. In particular, Figure 17 shows how the core literature has a high betweenness and PageRank respectively, meaning that it bridges different areas of the network. But the core also has a lower local clustering coefficient, due to the fact that it helps connect groups of sources which are more densely connected within the group but not across groups. The intuitive explanation is that groups of sources here represent the reference lists of a few citing books, which are fully connected among themselves but are only connected with other groups of such a kind through the core literature.



(a) The core literature (red/dark grey) and the rest (cyan/light grey).

(b) Betweenness centrality is higher for darker nodes (i.e. mostly the core).

(c) PageRank is higher for darker nodes (i.e. mostly the core).

(d) Local clustering coefficient is higher for darker nodes (i.e. not the core).

Figure 17: The core literature highlighted in the giant component of the co-citation network, the betweenness and PageRank centralities which are higher for the core, and the local clustering which is instead lower for the core. This visualization uses Gephi's Force Atlas 2 with LinLog mode and edge weight of 3.5.

Visual intuitions find confirmation using correlation coefficients, as shown in Table 7. Perhaps interestingly, and despite the fact that the core behaves as expected, the correlation coefficients are not as high as to warrant too narrow an explanation. The number of received citations in the directed network certainly determines the important role of the core into bridging groups of literature otherwise barely connected, but this role is not accounted for exclusively by the core. The core likely plays the prominent role in this respect, but other works too help in keeping the network connected. It should appear clear by now how using a threshold on the number of received citations is but one method to define the core literature. It could also have been individuated, with similar but not identical results, using the properties of the co-citation network, e.g. according to some centrality measures such as PageRank or betweenness. This was indeed one of the purposes for the introduction of co-citation networks in the first place [Small, 1973]. Different aspects of the core literature can, in this way, be put into play, besides its popularity (number of received citations).

Table 7: Pearson correlation coefficients among different measures and the core literature. All measures account for edge weights. *Is core* is a boolean field indicating if a node belongs to the core (1) or not (0).

Measure	**Is core**	**Degree**	**Betweenness**	**PageRank**	**Local clustering**
Is core	1	0.63	0.42	0.62	-0.41
Degree	0.63	1	0.82	0.99	-0.35
Betweenness	0.42	0.82	1	0.89	-0.26
PageRank	0.62	0.99	0.89	1	-0.36
Local clustering	-0.41	-0.35	-0.26	-0.36	1

Yet the main point holds: the core literature exists, and it is the main reason for which the area appears to be connected at the citation level. Scholars from different sub-areas and dealing with a variety of topics, still share a (small) set of works which they all refer to. The next section explores these works in more detail.

4.4 The core literature

The core literature, composed by 769 books cited ten or more times, is almost uniformly spread across periods of publication of the cited material. Still, it is older than the average due to the time needed to accumulate citations in this

Table 8: Summary of the two groupings of core sources, by age and by typology. Proportion indicates how many works per category are core, and is given in % over the three periods pre-1800, 1800-1949 and 1950-present. N. Citing indicates the number of citing books, from which citations were extracted, which also end up in a given category.

Group	**Number**	**Proportion**	**N. Citing**
by Age			
pre 1800	43	1.4	0
1800-1949	249	2.4	4
1950 to present	477	2.2	88
by Type			
Primary sources	107	-	1
Reference works	77	-	2
Scholarly monographs	585	-	89

research area. It also is, as a consequence, quite varied in its contents. Two groupings can be proposed for the core literature: one, more trivial, where core works are grouped by their publication age: pre-1800, 1800-1949 and 1950 to the present. Another perspective uses the typology of the publication itself, allowing to individuate three different groups: primary sources, works of reference and scholarly monographs. A summary is given in Table 8.

The first group of core works by age (defined as *age 1*) is composed of publications dating before the year 1800, mostly early printed books. Yet several of the most cited primary sources have been edited at a later time in a critical edition, made in order to provide easier access to historians. A notable example of primary source which was edited and published at a later time are the Diaries of Marin Sanudo, a Venetian nobleman who recorded the daily life of the city for several decades across the fourteenth and fifteenth centuries. This edition was published in between the years 1879 and 1903. Conversely, early works of scholarship published before the nineteenth century are also included in this category. A second group by age (*age 2*) is composed of sources published during the period between 1800 and 1949. This time in the historiography on Venice, developing since the fall of the Republic, is characterized by the efforts of local historians to cast a positive view on the city's past, but more importantly by the effects of the general positivistic turn in historical studies, which fostered the production of works of reference and overarching syntheses of the history of the Republic [Infelise, 2002; Povolo, 2002; Dursteler, 2013]. Works of reference can be critical editions of documents, with associated historical studies, as well as historical dictionaries, repertories, bibliographies or any kind of work meant to aid future historians by providing digested information. The most notable example is perhaps "Delle Iscrizioni Veneziane", by Emmanuele Cicogna, a wide repertory of Venetian epigraphs. Additionally, during the same period, modern historiography developed while ambitious works of historical synthesis were produced on the basis of newly discovered documentary evidence. An example is the Documented History of Venice by Samuele Romanin, published between 1853 and 1861. Several works in this group are multi-volume. A third and last group (*age 3*) is more recent and abundant, gathering all works published from the year 1950, in what we might term the contemporary historiography on Venice. This group of 477 books comprises some works of enduring importance such as the History of the Population of Venice by Daniele Beltrami (1954) or the Economic History of Venice by Gino Luzzatto (1961), but fewer works of reference and edition of sources. Every core group by age includes in between 1.4 and 2.4% of the cited works for the given period, with proportionally more works from period two being core than the other periods.

The groups by typology are organized differently. A first typology (*type 1*) comprises primary sources individuated by being publications or documentary records not originally meant as scholarly works, including critical editions. In practice, all non-scholarly publications plus all editions of documents are included in typology one. The third typology (*type 3*) comprises all works of scholarship, published at any time. Using this definition, several works from age one and, even more, age two, end up in typology three. Lastly, the second typology (*type 2*) gathers all works of reference made by historians for historians (for example catalogs, dictionaries, bibliographies, indexes and guides), according to the definition given previously. Most of these works have been published during the nineteenth and early twentieth centuries. A summary of this second classification method is as well given in Table 8, whilst the five most cited works per typology, along with their citation counts are further detailed in Table 9.

The presence of a core literature, and its three main typologies of primary sources, works of reference and scholarly works, highlights what connects the field. Notorious primary sources can become commonplace among historians. Works of reference often entail an investment of resources which is not easily replicated, thus determining their enduring importance. Some might even contain materials on long-disappeared records or artefacts, for which they represent the only surviving evidence. Works of reference are also often a product of specific periods during which their status as a scholarly product was deemed on par, if not above that of scholarly monographs, such as during the second half of the nineteenth century. Primary sources and works of reference can be considered as shared for the community, works on top of which it is possible to build further scholarship, and that do not fall into oblivion until another comparable and better work is acknowledged in their place. Lastly, scholarly monographs of recognized status emerge quite slowly, often after one or more generations have passed. Clearly, citations in the humanities accumulate at a slow pace, especially so for books. Yet the fact that recent historiography so often cites old scholarship can be explained in several ways: for once, topics long forgotten can live through a second life, such is the case for private life and the history of interiors, a topic early discussed by Pompeo Molmenti in his highly cited work (the most cited in typology three), and rediscovered by several scholars since thirty years ago. Another motivation to cite old, well-known works is that they are, effectively, widely recognized, thus mentioning them is important to signal membership in the community. The importance of citing to contextualize or signal, especially in books where citations are more abundant, might be a factor contributing to the importance of the core literature. Lastly, highly cited are also landmark works which originated, or anyway highly contributed to a specific topic of enduring relevance, thus they are cited in order to reconstruct its main developments.

By considering the use of the core literature over time, in Figure 18, it is shown that the proportion of citations to the core literature is relatively stable over different typologies. Typology one and two comprise in fact fairly specialized works, which are marginal in terms of the total number of received citations, but stable in their presence. Typology three is instead more substantially represented, rising and leveling-off at 20% received citations over recent decades. With respect to the categories of core literature by age, it is possible to appreciate the waning-out of older scholarly literature being displaced by more recent works over time, in a process of slow update of the scholarly literature of reference which does not impact primary sources nor works of reference. The proportion of references to old literature is in fact slightly rising over time. We consider (a modified version of) the Price index [De Solla Price, 1970], or the proportion of citations to works published

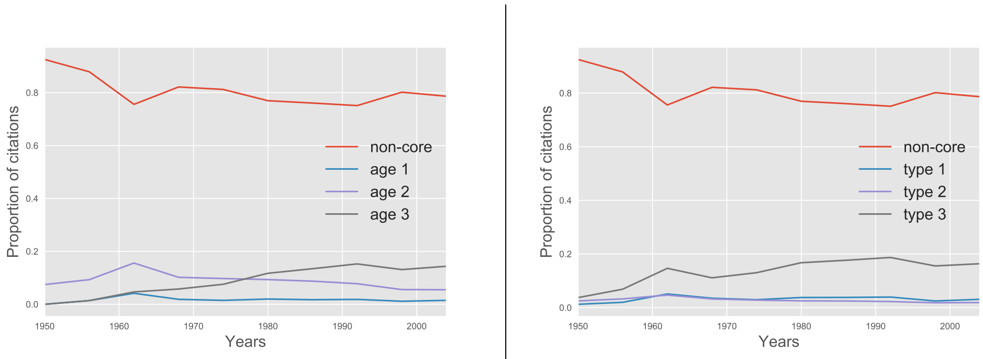
Table 9: The top core works by typology. Multiple editions of the same work are found sometimes in the dataset; if that was the case the number of citations to the most cited edition is given first, and the total number of citations to the work follows in parentheses.

Title	**Author**	**Year**	**Citations**
Core by typology, cat. 1			
Venetia, cittá nobilissima et singolare	Francesco Sansovino	1581, 1663 (1998)	90 (291)
Ecclesiae Venetae antiquis monumentis	Flaminio Correr	1749, 1758, ..	116 (198)
Delle memorie venete antiche profane ed ecclesiastiche	Giambattista Gallicciolli	1795	93 (110)
I diarii	Marin Sanudo	1496-1533 (1879-1903)	38 (79)
De origine, situ et magistratibus urbis Venetae, ovvero La città di Venetia	Marin Sanudo	1980	72 (77)
Core by typology, cat. 2			
Delle inscrizioni veneziane	Emmanuele Cicogna	1824-1853	139 (177)
Dizionario del dialetto veneziano	Giuseppe Boerio	1829, 1856, . . .	67 (113)
Saggio di bibliografia veneziana	Emmanuele Cicogna	1847	64 (69)
Dizionario del diritto comune e veneto	Marco Ferro	1845	60 (63)
L'Archivio di Stato di Venezia: indice generale, storico, descrittivo ed analitico	Andrea Da Mosto	1937+	43 (45)
Core by typology, cat. 3			
La storia di Venezia nella vita privata dalle origini alla caduta della Repubblica	Pompeo Molmenti	1880, ...	87 (222)
Storia documentata di Venezia	Samuele Romanin	1853-1861	122 (163)
Storia economica di Venezia	Gino Luzzatto	1961	75 (112)
Storia della popolazione di Venezia	Daniele Beltrami	1954	93
Rich and poor in Renaissance Venice	Brian Pullan	1971	74 (91)

maximum 10 years before the citing one, in Table 10. The values, already very low, are slowly decreasing over time. This is interesting as it points to a possible growing preference of scholars for older and well-known sources, instead of more recent (and abundant) literature. As the core literature is often old, this would mean that its relative importance is slowly growing over time. The top-cited sources over the same intervals of time highlight in fact the stable popularity of core sources of typology one and two, with some change happening in typology three (results are omitted here for brevity, and can be found in the Chapter's code repository).

Table 10: 10-year Price Index over intervals of citing sources. The Price Index is the proportion of citations to works published maximum 10 years before the citing one.

Period	**Mean**	**Median**
Until 1980	0.235	0.21
1980-1990	0.221	0.21
1991-2001	0.216	0.205
2002-2014	0.2	0.194



(a) Proportion of citations to the core by age.

(b) Proportion of citations to the core by typology.

Figure 18: The proportion of citations given to core and non-core works over time. Proportions are calculated using a smoothing window of six years, for every point in time the total (y axis) sums to one. The proportion of citations to age category two reduces, and category three rises, as new scholarship supplements older works in recent years. With respect to typologies instead, we see that the role of typologies one and two is marginal but stable, whilst typology three rose to occupy a stable 20% of citations which are given to highly cited, well-known scholarly monographs.

The proportion of citations to the core literature can be compared with the proportion of citations given to uniquely cited works, or works that are cited only once in the dataset. These distributions are given jointly in Figure 19. Interestingly, citing books have a more uniform distribution of citations to unique works, with a mean to 30-40% but high variance, whilst core works occupy a more limited yet significant role, taking on average 10-20% of citations. Most books balance their citations to a fraction of core works and less well-cited works, in what appears to be a trade-off between contextualized and specialized referencing.

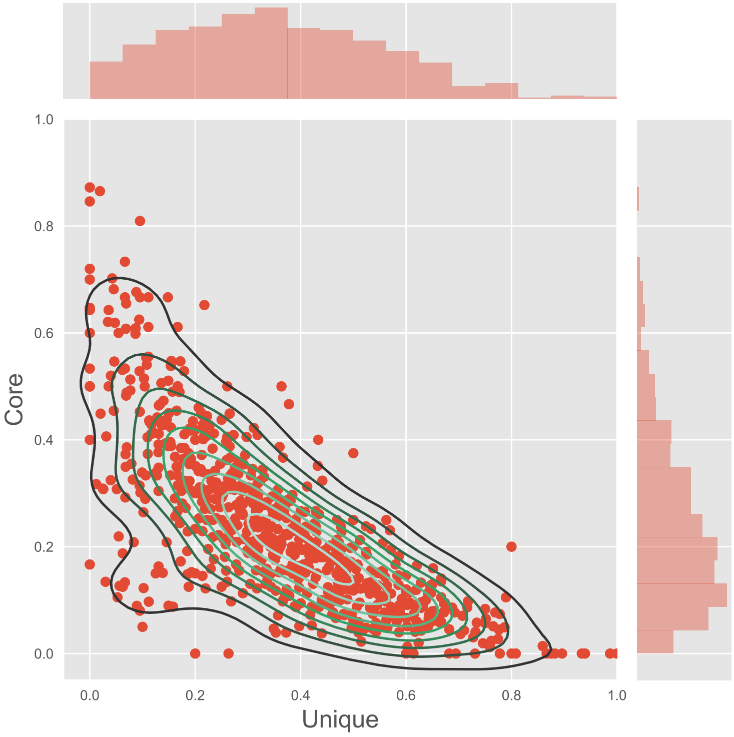


Figure 19: The joint distribution of the proportion of citations given to the core literature and to uniquely cited works (i.e. works cited only once). Most citing books cite 10-30% core and 30-50% uniquely cited works.

The core literature, which glues together the field of the history of Venice, represents all periods of its development, as well as different typologies of publications. The historians of Venice share, it seems, a set of sources, works of reference and monographs which are widely known by practitioners, and remain relevant to this day of a rapidly increasing variety of perspectives [Grubb, 1986; Horodowich, 2004]. A limited set of well-known books which accrue sufficient recognition to become cited even outside of their original specialization, and part of the common ground of the scholars of the field. On one side we have primary sources and works of reference, which never become outdated until substituted, on the other scholarly works of particular importance, which are slowly updated, or rediscovered over time, as the field shifts attention to different topics, but grounds them in previous work. This situation might well be shared in other research areas in history and beyond.

4.5 Discussion

This chapter explored and highlighted the importance of the core literature in bridging different clusters of publications into a coherent area of research. We focused on book to book citations where source items (citing works) were selected from recent literature on the topic of the history of Venice. A fine-grained manual classification was used in order to qualify the results of different clustering methods. Book citations individuate what is likely the most general and encompassing citation level in the humanities, thus the assumption was that they would yield the most coarse organization of the considered community. It is shown that the historians of Venice broadly organize by discipline and historical period at this citation level. The community's connectivity is strongly reliant on few, core sources. Core sources divide into three categories in turn: highly cited monographs, reference works and (edited) primary sources. The reliance on the core literature is also found to be increasing over time, meaning fewer and fewer of them are shared across clusters.

We therefore confirmed that the historiography on Venice presents a holistic intellectual organization. A group of core, highly-cited works is the main motivation for which both the recent literature (bibliographic coupling network) and the intellectual base (co-citation network) are almost connected, and organized in coherent clusters bridged by it. The structural reliance on the core literature is also found to be rising over time, as the field becomes increasingly more varied. The core literature mainly comprises primary sources and works of reference, which never age out until substituted by similar contributions, and represent the scholarly contributions on *evidence* and means to access it more effectively. The core literature also includes scholarly monographs of substantial importance, which become well-known within the community, and constitute scholarly contributions of *interpretation.* This second group of core works is instead slowly updated over time, as the area of research moves to new topics or casts new light on older ones. Despite the fact that the humanities and social sciences will likely never become high-consensus, rapid-discovery sciences, the role of some primary sources and works of reference in order to ground their discussions can perhaps be compared to the 'genealogies of research technologies' so important to allow for the cumulative advance of the sciences [Collins, 1994]. Their impact over time is perhaps a still under-acknowledged element with respect to the evaluation of research in the humanities, and one which directly speaks to the possible contribution of the Digital Humanities in the long run.

Interestingly, in the case of Venice, an established tradition of studies and resources still bears an influence on recent scholarship, which is growing considerably more elaborated and internationally oriented. The presence of a core literature is ultimately the reason for which we can still consider the historiography on Venice a field of research on its own, instead of a set of increasingly fragmented areas. At this point, we can tentatively advance two more general considerations. Firstly, that the core literature can influence a community for a very long time. This has implications for research evaluation, which evidently cannot be based on short-term citation counts. Secondly, that the pace of research in recent times is likely resulting in intellectual fragmentation, as the rising importance of the core suggests. This trend could possibly be attenuated by increased efforts to produce and use shared (digital) resources and reference works, which have the proven long-lasting effect of integrating research efforts.

In conclusion, this first mapping effort highlights what is perhaps an only superficial well-connectedness of a community of historians. In fact, it seems that even at what we might regard as the broadest citation level possible, that of books, only few core sources are shared within sub-communities of historians, and increasingly less so with time. This result thus prompts a set of further questions: how does the core literature behave structurally, more in general? What is in turn the structural role of different publication typologies such as journal articles and books? Are there general trends, such as towards increasing intellectual fragmentation (i.e. reliance on fewer core sources for general connectivity)? Lastly, how does the intellectual organization of historians compare with other disciplines in the humanities and beyond? The following chapters will consider these questions in order, so to sketch a comprehensive view on the intellectual organization of history.

5 The structural role of the core literature

The secondary literature of historians takes a variety of forms and covers a multiplicity of intentions. Hicks [2004] individuates four literatures in the social sciences, relevant for most disciplines in the humanities too: international journal articles, books, national literature (whose scope is local due to its topic and citation span, not necessarily its language) and non-scholarly publications. International journal articles, the main focus of commercial citation indexes, are likely just the tip of the iceberg. Books, albeit fewer in number, have a disproportionate impact in terms of received citations, and manifest specific citation behavior patterns [Thompson, 2002]. Furthermore, books are of importance as they make for the best part of the core literature in a field, as discussed in Section 2.3 and shown in Chapter 4 with respect to the historiography on Venice. All these publication typologies serve complementary purposes. In fact, if we follow Nederhof [2006] and assume that there are three target audiences for the social sciences and humanities, we have: other scholars at the research frontier (assuming this group includes somewhat internationally recognized scholars); regional or national scholars (considering scholars mostly dealing with national literature in the sense of Hicks [2004])[18]; and the non-scholarly public.

Taking the somewhat wider perspective of what can be cited by such literature, we find variety in abundance. A non-exhaustive list includes: primary sources (surely a vast category in itself), books, journal articles, conference proceedings and contributions in edited volumes, works of reference and edition of sources, databases and online resources, book reviews, plus all kind of writing for the general public such as other books, essays, newspaper or online articles and even blog posts. We should thus abandon any monodimensional view of humanities' scholars and the intellectual landscapes they inhabit [Watson-Boone, 1994; Larivière et al., 2006b]. More likely, several profiles of scholars in the humanities exist, each having a tendency for using a combination of the aforementioned typologies of sources and publications[19]. Any effort to map the humanities should thus acknowledge their *multidimensional intellectual organization*, and the fact that using only a few of these

[18] Admittedly, the dichotomy international/frontier and, by implication, national or regional/not frontier, is probably too simplistic.

[19] See e.g. Verleysen and Weeren [2016a,b] and Emmeche et al. [2016], especially Svend Østergaard and Peter Lau Torst Nielsen, *Research Styles: Data and Perspectives in the Human Sciences* therein.

sources will inevitably lead to a simplified view of their complexity.

We consider in this chapter two related questions. Firstly, can we say that a core literature exists in historiography, as is the case for that on Venice? In particular, we would like to know if a core literature exists for different citing publication typologies, and what it comprises. Secondly, what structural role does the core literature play with respect to the intellectual landscape defined by citing publications? More precisely, is the core literature spanning only specific locations, or does it connect far-apart areas of the landscape? Are different core sources, such as books and journal articles, behaving differently in this respect? In order to scale our analysis we will here consider results from three citation datasets. First, the book to book dataset on the history of Venice explored in Chapter 4; secondly, all Web of Science (WoS)-indexed articles published in *The Library*, a renown journal in the area of the history of the book, with all WoS source and non-source items they cite; lastly, all historiography from WoS with all cited source items. As before, in what follows we refer to books generally to indicate a variety of publications that take this form.

In summary, this chapter focuses on the following:

1. Define a general method to assess the structural role of cited publications into connecting the resulting bibliographic coupling network of an area of research, via a set of complementary indicators.

2. Use the indicators to analyze the core literature in different datasets on history, showing that a core exists and is mainly composed of books, where applicable. This result allows us to generalize some of the outcomes of Chapter 4.

3. Further advance our understanding on what role the core literature plays, by showing that it can act both globally and (more often) locally, and that only few sources, mostly books, provide global connectivity.

This chapter is based on published work in Colavizza [2017b].

5.1 Method

Given a set of citing publications, representing an area of research of interest, its core literature can be defined in a variety of ways. We consider here as core the most cited publications in a given dataset. Such core literature can then be analyzed with respect to the bibliographic coupling network of the citing publications, which defines the intellectual landscape of the area of research. More specifically, with respect to a partition of such network into communities. Given that this network's structure necessarily relies to a considerable degree on the core literature, our goal is to qualify different roles that core sources can assume with respect to the communities of the bibliographic network. Four indicators, defined for a core source c, are introduced:

- *Within indicator* a_c: captures the importance of the source c to connect citing publications within the same communities.
- *Between indicator* b_c: captures the importance of the source c to connect citing publications across communities.
- *Topicality indicator* c_c: captures the relative importance of the source c to connect citing publications within a specific community or within several communities. Topicality quantifies how focused the action of a source c is within a specific community.
- *Bridging indicator* d_c: captures the relative importance of the source c to connect citing publications between a specific pair of communities or between several pairs. Bridging quantifies how focused the action of a source c is across a specific pair of communities.

The four indicators capture different aspects of the role of the core literature with respect to relations between citing publications. They indicate how important the core source is to connect communities internally (within) or among each other (between), and how focused this action is (most influence within one community or between a pair of communities).

More formally, we start with the following setting. Take $D = (V_D, E_D)$ the directed citation network of the set of publications under consideration, where vertices $v \in V_D$ are both citing publications and cited sources, and $e(v_1, v_2) \in E_D$ are directed edges between such vertices. A source can be both citing and cited, thus D is not in principle acyclic. Take the projection of D onto citing publications $B = (V_B, E_B)$: the weighted bibliographic coupling network. B can also be represented by its square and symmetric weighted adjacency matrix W. For simplicity, without loss of generality, we consider raw weights of one for each reference in common between any two citing publications. Take L, a partition of B into communities, where every vertex $v \in V_B$ is assigned to a single community. Lastly, take a set of core sources $C \subseteq V_D$. Core sources can be individuated in a variety of ways, for example by taking a certain top quantile of the in-degree distribution (number of received citations) of D. To be sure, any cited source in V_D can be considered for analysis, core sources being a particularly interesting subset.

All indicators are based on the idea of considering the contribution of a source in the core C to the weight of the edges in B, the bibliographic coupling network, taking into consideration its partition into communities. Consider the function:

$$cit(i,j,c) = \begin{cases} 1 & \text{if } \exists e(i,c) \in E_D \wedge \exists e(j,c) \in E_D \\ 0 & \text{otherwise} \end{cases}$$

That is to say, if both i and j cite c in D, $cit(i,j,c)$ returns 1, which is the weight contributed by c in the edge between i and j in B. This function assumes raw weights were used to construct the bibliographic coupling network. That is to say, $W_{i,j} = \sum_{v \in V_D} cit(i,j,v)$. Other weighting schemes might be used, such as fractional counting [Perianes-Rodriguez et al., 2016], and then cit shall be modified accordingly.

We can now proceed to establish a preliminary version of our indicators, defined for every $c \in C$ as follows:

$$\alpha_c = \sum_{l \in L} \alpha_c(l) = \sum_{l \in L} \sum_{i,j \in V_B} cit(i,j,c)\delta_l(i,j) \tag{1}$$

$$\beta_c = \frac{1}{2} \sum_{l \in L} \beta_c(l) = \frac{1}{2} \sum_{l \in L} \sum_{i,j \in V_B} cit(i,j,c)(1 - \delta_l(i,j)) \tag{2}$$

$$\gamma_c = \max_l \alpha_c(l) \tag{3}$$

$$\delta_c = \frac{1}{2} \max_l \beta_c(l) \tag{4}$$

Where l_i is the community to which i is assigned according to partition L, and $\delta_l(i,j) = 1$ if $l_i = l_j$, 0 otherwise. Note that $\alpha_c + \beta_c = \sum_{i,j} cit(i,j,c)$, the total edge weight contributed by c in B. γ_c and δ_c only consider the community l yielding the maximum contribution to α_c and β_c respectively.

The division by $1/2$ is needed in Equations 2 and 4 since when i and j belong to different communities, the contribution of c is considered twice.

Yet the degree distribution of citation networks is commonly skewed, thus the core has by definition a disproportionate role in the structure of the bibliographic coupling network. The problem with these indicators is that they do not account for the obvious effect of the in-degree of core sources. We would like, instead, to be able to compare different core sources, and core sources from different datasets, irrespective of the underlying degree distribution. As a consequence, we need a null model to compare against. A valid choice is the *configuration model* (cf. Newman [2010]; Barabási [2016]). In a directed setting, having the list of vertex pairs in two arrays (citing and cited respectively) of equal length, an instantiation of the configuration model consists of randomly permuting one of the two arrays, to produce a random network with the same degree distributions (both in and out degrees) as the original one. A minor adaptation is the need to "simplify" the so-created random network by removing eventual self-loops and multi-edges (low-probability events in themselves). Such network can serve as a null model to test some properties of the original network, disregarding the effect of its degree distribution. It is good practice to produce several instantiations of such configuration model and average out the desired statistics.

In our case, we take N instantiations of the configuration model of the directed network D, each time construct a new bibliographic coupling network and calculate as follows:

$$\chi_c = \sum_n^N \sum_{l \in L} \sum_{i,j \in V_B} cit^n(i,j,c)\delta_l(i,j) = \sum_n^N \chi_c^n(l)$$

$$\phi_c = \frac{1}{2}\sum_n^N \sum_{l \in L} \sum_{i,j \in V_B} cit^n(i,j,c)(1-\delta_l(i,j)) = \frac{1}{2}\sum_n^N \phi_c^n(l)$$

$$\psi_c = \sum_n^N \max_l \chi_c^n(l)$$

$$\omega_c = \frac{1}{2}\sum_n^N \max_l \phi_c^n(l)$$

Where cit^n considers edges in the nth instantiation of the configuration model. Note that we keep the same partition L at all times. The final indicators are:

$$a_c = \frac{\alpha_c}{\alpha_c + \beta_c} - \frac{\chi_c}{\chi_c + \phi_c} \tag{5}$$

$$b_c = \frac{\beta_c}{\alpha_c + \beta_c} - \frac{\phi_c}{\chi_c + \phi_c} = -a_c \tag{6}$$

$$c_c = \frac{\gamma_c}{a_c} - \frac{\psi_c}{\chi_c} \tag{7}$$

$$d_c = \frac{\delta_c}{b_c} - \frac{\omega_c}{\phi_c} \tag{8}$$

The behavior of the indicators is as follows: a_c is positive if the core source is more important than its degree would justify in connecting nodes within the same communities, b_c if across communities. The more positive c_c is, the more the action of c in connecting nodes happens within the same community, the more positive d_c is, the more the action happens between the same pair of communities, irrespective of the effect of the in-degree of c. Note that a_c is in general positively correlated with the modularity of the partition L [Newman and Girvan, 2004], and b_c negatively so. As a consequence, any value of these two indicators is related to a given partition of the nodes. If the partition in use is the result of a stochastic modularity maximization procedure, all indicators should thus be averaged over multiple partitions. We eventually note in passing that indicators c_c and d_c could have been expressed in other ways too, for example by considering the entropy of the distribution of the contribution of c within or between all possible communities or pairs of communities.

The relation of a_c (and b_c) to the modularity of the partition L is of interest. Modularity measures the density of the links within communities, as established by L, by comparing it to a random null model. In our setting, the (weighted) modularity Q of L is defined as:

$$Q = \frac{1}{2w} \sum_{i,j \in V_B} \left[W_{i,j} - \frac{W_{i,*} W_{j,*}}{2w} \right] \delta(l_i, l_j)$$

Where w is the sum of the weights of all edges in W, the weighted adjacency matrix of B, and $W_{i,*} = \sum_j W_{i,j}$, the weighted degree of vertex i. If we take the perspective of c, our core source of interest, we can construct a bibliographic coupling network $B^c = (V_B, E^c_B)$, and related W^c, only on the basis of the edges established by coupling citations to c. Note that the adjacency matrix W^c is binary at this point. In this setting, an alternative definition to the within indicator (and the between, similarly) can be based on modularity as follows:

$$a_c^* = \frac{1}{2w^c} \sum_{i,j \in V_B, \exists e(i,j) \in E_B^c} \left[W_{i,j}^c - \frac{W_{i,*} W_{j,*}}{2w} \right] \delta(l_i, l_j)$$

That is the modularity of L considered only on the weights contributed by c, whose total sum is w^c. The main difference between a_c and a_c^* rests in the use of different null models, calculated over different networks: the configuration model establishes random multigraphs with a given degree sequence, here over D, the directed network; the null model used in modularity, called the Chung-Lu model, establishes random simple graphs with a given degree distribution, here over B, the bibliographic coupling network. The modularity Q of L given above is an aggregated function of this last alternative indicator, therefore the general distribution of both the within and between indicators are influenced by it. Nevertheless, individual core sources can behave in a variety of ways under this general setting.

5.2 Datasets

We use three datasets, motivated by the desire to consider the role of different core literature and citing publications with respect to their publication typology. The main difference cast here relates to the distinction between books and journal articles, both as citing and cited sources. In order to attempt such a comparison, we have to consider quite different datasets: a) the book to book citations in the sub-field of the history of Venice, introduced in Chapter 4; b) a dataset of article to both book and article citations, extracted from a specific journal in the sub-field of the history of the book, called *The Library.* We considered its WoS-indexed articles and what they cite, either source and non-source in WoS; c) a third dataset of article to article citations, from all WoS subject “History”, limiting cited sources to what is indexed in WoS. These datasets should allow us to discuss sets of core literature composed by books cited by books, articles and books cited by articles, and articles cited by articles, yet it must be stressed that they are compared not as equals, but as a means to explore the same phenomenon from different angles, as it is made possible by the (limited) availability of data. A summary of the three datasets is given in Table 11. There is a disparity in terms of size and coverage among these datasets, largely due to data availability. This especially entails the fact that dataset three, the largest one, should be considered somewhat apart of the other two, as will be discussed in what follows.

We identify the core literature in each dataset by taking all sources in the top 99.5 percentile of the in-degree distribution (number of received citations) of every directed graph. All in-degree distributions are skewed (omitted here) and highlight a few, highly cited sources.

Table 11: Summary statistics for the three datasets under consideration. The threshold of the number of received citations at the 99.5 quantile, used to establish the set of core sources, is given in the last row.

Statistic/Dataset	**Monographs History Venice**	**The Library 1981-2016**	**All History 2005-2015**
Citing typology	Monographs	Journal articles	Journal articles
Cited typology	Books	Books and articles	Journal articles
# citing publications	673	479	36'709
# cited sources	36'922	11'237	101'777
# edges	68'525	13'176	159'610
# core sources (99.5 quantile)	129 (22)	65 (6)	776 (9)

In order to find the partitions of the bibliographic coupling networks we use a modularity maximization approach [Newman and Girvan, 2004], in the popular Louvain implementation [Blondel et al., 2008]. Despite its shortcomings, this method produces high quality results and is widely known in both the networks [Fortunato and Hric, 2016] and bibliometrics communities [Šubelj et al., 2016]. All indicators were calculated averaging results from ten possible partitions made using the Louvain algorithm with default resolution parameter at one, and for each one a hundred instantiation of the configuration model were averaged.[20] Modularity maximization tends to produce larger communities with larger datasets, when using the same resolution parameter. This is important to keep in mind for dataset three, which is larger in size than datasets one and two.

[20] Analyses relied on igraph [0.7.1] [Csardi and Nepusz, 2006] and Vincent Traag's community detection library [0.5.3] available at `https://github.com/vtraag/louvain-igraph`.

5.2.1 Books of the historians on Venice

The first dataset considers the specific sub-field of the history of Venice. This dataset comprises relatively recent books in a variety of languages, selected through library resources such as catalogs and shelving strategies. The procedure followed to extract their citation data is detailed in Chapter 3. The dataset is freely available online in multiple formats [Romanello and Colavizza, 2017]. This dataset comprises 673 citing books, which cite 36'922 books in turn, with 68'525 unique citations.

As discussed at length in Chapter 4, this dataset of citing books comprises works from different communities: medieval, early modern and modern history, art history and history of architecture, plus a variety of specialties therein, such as economic or gender history. The core literature of this dataset is mainly composed of primary sources, works of reference and renown scholarly books. Primary sources are often edited documents (e.g. the diaries of Marin Sanudo) or early printed works. Reference works such as repertories, inventories and dictionaries, often product of the local historians of the 19^{th} century, are still largely in use today. Lastly, scholarly monographs of lasting importance include some works from the 19^{th} century, as well as more recent literature published since the 1950s.

5.2.2 The Library: articles of the historians of the book

Our second dataset considers a different field of history: the history of the book, and a different publication typology: journal articles. We consider one of the most renown journals in this context: *The Library: Transactions of the Bibliographical Society* of London. Historians of the book extensively rely on resources such as catalogs and repertories for their work, they are organized into quite specialized communities with strong bonds with other fields such as library and information science, literature and philology.

All the indexed research articles in WoS are considered, from 1981 to 2016 included, for a total of 491 articles. We consider both source and non-source items by exporting all references from the WoS interface. Exported articles were processed using the Sci2 tool [1.2 beta] [Sci2 Team, 2009], in order to extract the directed citation network. Sci2 allows to detect duplicate nodes (i.e. references pointing to the same item) by comparing all references using the Jaro-Winkler measure.[21] Groups of references similar above a certain

[21] See `http://wiki.cns.iu.edu/display/CISHELL/Detect+Duplicate+Nodes`, ac-

threshold are retained as candidates for merging. This method is far from perfect, but allows to create a set of grouped references to be manually checked for refinement. We proceeded as follows: first, all references with no author were removed from the dataset, as too problematic to disambiguate. This is the case, for example, for references to newspaper articles. Secondly, all paginations were removed, given that they often pointed to the specific location which was cited (as it is practice in the humanities) instead of the page interval of the cited article (as in the sciences). The Sci2 tool was then used to detect groups of references to be merged, with a threshold of 0.84 on the Jaro-Winkler measure, established empirically by finding the threshold, rounded at two digits, which would yield a precision of less than 0.5 in the 100 pairs of references to be merged with a similarity just below that threshold. On this dataset, a threshold of 0.84 had a precision so calculated of 0.41. Note that we left inevitably out some references as not grouped, but the number of false negatives decreased rapidly thereafter. Lastly, all retained groups of references to be merged were manually checked and cleaned. During cleaning, multiple editions of the same work were considered as one. The result is that 479 articles cite 11'237 unique items. Some citing articles are removed because they did not possess extracted references (9) or they were merged (2). The number of references in the original dataset is 14'412, the number of citations after clean-up is 13'180.

The core literature of this dataset is mainly composed of primary sources, works of reference and seminal monographs. Examples include the overly important Shakespeare, the records of the Stationers Company and early printing manuals such as Moxon's "Mechanick Exercises"; catalogs (e.g. the "English Short Title catalog"), dictionaries or reference works (e.g. Plomer's "Dictionary of Booksellers and Printers"); renown monographs (such as Gaskell's "New Introduction to Bibliography"). Only 13 of the 100 most cited sources are journal articles. The core literature resembles the one from Venice in its assortment, with quite more emphasis on reference works: the cornerstone of studies in bibliography and the history of the book.

The main intellectual communities of this dataset are shown in Figure 20. The strong focus on English studies, and the interconnection of the history of the book and literature clearly emerges, in particular for Shakespeare studies. Topics from continental book history seem marginal instead. Yet the community publishing in *The Library* appears well organized in specific

cessed May 2017.

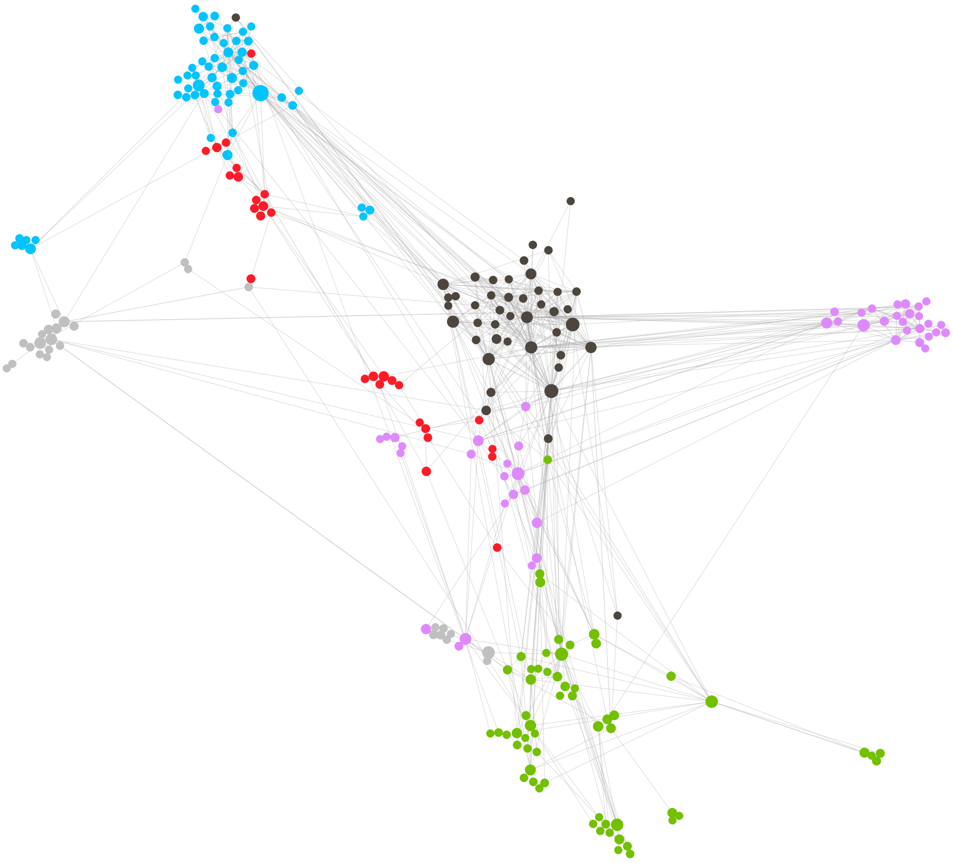


Figure 20: The communities of *The Library* dataset, using the Louvain method with configuration identical to experiments. This network has been trimmed from edges of weight less than 2, as a consequence nearly 60% nodes are visible. The communities are: cyan – early English printing; pink – Shakespeare studies; green – English literature 16-17th century; gray – Renaissance book production in the European continent; dark gray – English book production and commerce in early modern times; red – Libraries and collections. The size of the nodes is proportional to their betweenness centrality. This visualization was made with Gephi 0.9.1 [Bastian et al., 2009], using Force Atlas 2 with default parameters but for LinLog, dissuade hubs and prevent overlap modes active, scaling 2.0 and edge influence 1.5 [Jacomy et al., 2014].

areas of activity.

5.2.3 All of history in the Web of Science

The last datasets comprises all articles indexed in WoS under the WoS subject category of "History", published from 2005 to 2015 included. No other publication typology besides research articles was considered. With respect to their citations, everything that was indexed in WoS is retained, also outside of this specific subject category. Citations were taken from the CWTS databases [Olensky et al., 2016]. This dataset comprises 36'709 citing articles and 101'777 cited articles, with 159'610 citations among them.

The core literature of this dataset is composed mainly of seminal articles which delivered novel methods or arguments of enduring importance for a broad area of historical studies. Examples include gender history (Scott, 1986, *Gender, a useful category for historical analysis*), politicization of the past (Hall, 2005, *The long civil rights movement and the political uses of the past*), comparative history of development (Acemoglu, 2001, *The colonial origins of comparative development: An empirical investigation*) and cultures (Subrahnanyam, 1997, *Connected histories: Notes towards a reconfiguration of early modern Eurasia*), political history (Elliott, 1992, *A Europe of composite monarchies*).

The communities of this dataset are of more difficult evaluation, given the size of the network. Indeed, the size of these communities is also varied, as shown in Figure 21. For these reasons, and given the low coverage of WoS with respect to history, results from this third dataset are included only for reference, in order to highlight the structural properties of a larger, more sparsely connected network.

5.3 Results

The indicators applied on these datasets should allow us to highlight different properties of the core literature. We start with some hypotheses. First of all, that the core literature is present and plays a crucial role into connecting the bibliographic coupling network of the respective dataset. Secondly, that core books contribute more on average to the global connectivity of the network than core journal articles. If this were the case, their within indicator should be lower and their between indicator higher. Similarly, in such a case core books should be more likely to have a low bridging capacity, as they would

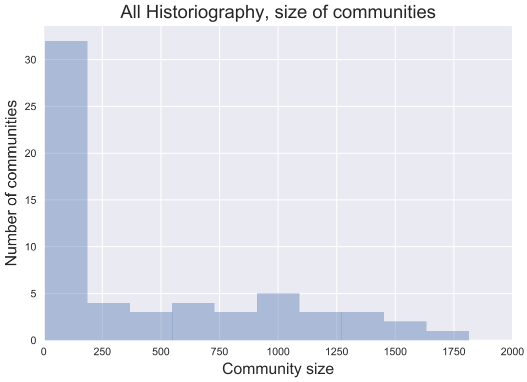


Figure 21: The distribution of the size of communities in the all history network, using the Louvain method with configuration identical to experiments. Shown are only communities with more than five articles. An inspection of the first six communities, by reading a random sample of 200 article titles each, led to the following broad classification (in order of decreasing size): Economic history; Intellectual and cultural history (pre-contemporary); Social history; Gender, slavery and minorities history; Colonial and post-colonial history; Contemporary political history.

not just connect two specific communities but several more. Lastly, we see no reason for their topicality to differ from that of journal articles, since it makes intuitive sense that any source is better known to a specific community than in general.

The starting point of the analysis are the bibliographic coupling networks of the three datasets. Table 12 reports their summary statistics. Most notably, the three networks differ in the basic terms of their connectivity. The first dataset, books on the history of Venice, results in a very dense and well-connected network; less so for *The Library*; and quite less so for the history dataset from WoS, which comprises a 30% of vertices in small components.

Table 12: Summary statistics of the bibliographic coupling networks.

Statistic/Dataset	**Books History Venice**	**The Library 1981-2016**	**All History 2005-2015**
# vertices	673	479	36'709
# edges	87'168	4435	161'802
# connected components	3	58	9901
Vertices in the giant component	99.7%	87.9%	69.5%
Network density	0.38548	0.03874	0.00024

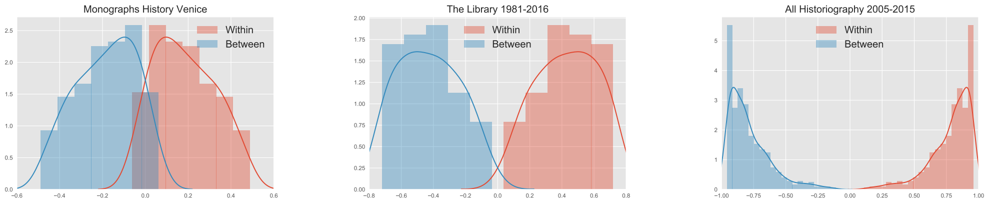
The statistics of indicators given in Table 13 confirm in part our initial hypotheses. The average modularity of partitions, influenced by the connectivity of the networks (higher connectivity usually implies a more difficult community partitioning task, thus lower resulting modularity), is much lower for the first dataset, and incrementally higher for the other two.

Table 13: Mean (median) value of the indicators over different datasets, plus the modularity of partitions. Values are averaged over ten partitions.

Statistic/Dataset	Books History Venice	The Library 1981-2016	All History 2005-2015
Within indicator	0.18 (0.17)	0.43 (0.43)	0.78 (0.83)
Between indicator	-0.18 (-0.17)	-0.43 (-0.43)	-0.78 (-0.83)
Topicality indicator	0.32 (0.37)	0.41 (0.44)	0.61 (0.62)
Bridging indicator	0.16 (0.14)	0.48 (0.48)	0.45 (0.43)
Modularity of partition	0.1835	0.4355	0.7135

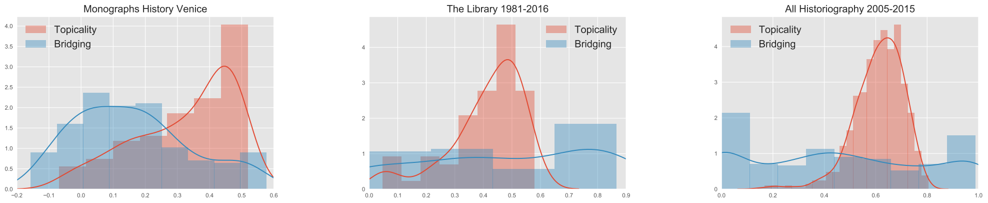
The within and between indicators behave as expected, their distributions are shown in Figure 22. Indeed, a core literature made only of books acts not just within specific communities but across them. If we consider the third dataset, the action of the core is considerably more limited to providing within-community connectivity. The second dataset, where the core literature is mixed, lays in-between these two extremes.

With respect to the topicality and bridging indicators, results are less unequivocal. Their distributions are given in Figure 23. If we can appreciate that indeed topicality is increasingly higher moving from dataset one to three, bridging captures a variety of behaviors, resulting in an almost uniform distribution of values. Core sources can thus bridge several or few pairs of communities irrespective of their typology. Nevertheless, there is a clear higher concentration of low-value bridging core sources in the first dataset, as shown in Figure 23a. The core literature from the books on Venice presents a higher proportion of low-bridging core sources, and a longer left tail in the topicality distribution as well. This entails that the connectivity action of these core sources is not just focused on a single community or community pair.



(a) Books History Venice. (b) *The Library*. (c) History WoS.

Figure 22: Distributions of the within and between indicators. Histograms are normalized, the lines are the kernel density estimations.

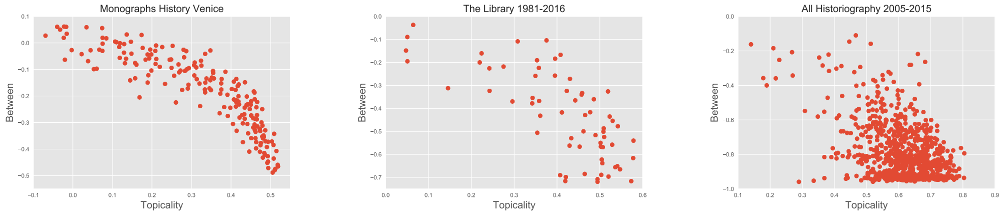


(a) Books History Venice. (b) *The Library*. (c) History WoS.

Figure 23: Distributions of the topicality and bridging indicators. Histograms are normalized, the lines are the kernel density estimations.

The more varied behavior of a core literature composed of books is highlighted in scatter plots which consider topicality and between indicators at the same time, in Figure 24. In dataset one, a clear pattern exists by which low-topicality sources have high between score, and vice versa, in two distinct linear regimes of change. Conversely, the norm for sources in the third dataset is to have very high topicality and varied, but comparatively lower between scores. The second dataset, as usual, presents mixed results.

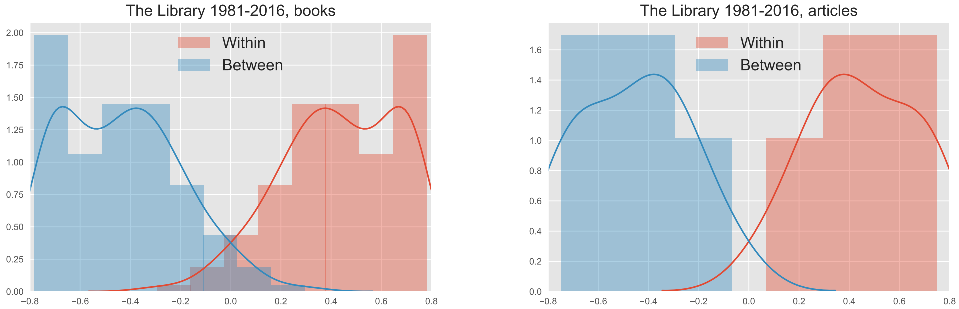
The implication is relatively clear: a core literature composed of books contains a higher proportion of globally interconnecting sources, spanning way outside their main community. Such segment of the core literature interconnects communities at a global structural level, where the action of another part of the core literature, and most journal articles in it, is structurally localized.



(a) Books History Venice. (b) *The Library*. (c) History WoS.

Figure 24: Scatter plots of the topicality and between indicators. Note that the within indicator is symmetric (positive) of between, around zero.

To further explore the difference between core books and journal articles, we consider dataset two and enlarge the number of core sources under consideration by taking the 99 quantile. There are 167 core sources now, of which 154 are books and still only 13 journal articles. The distribution of their within and between indicators, given in Figure 25, highlights the higher global reach of some books, being also negative within score, not to be found among journal articles. To be sure, most books still act locally as do most journal articles.



(a) *The Library*, books. (b) *The Library*, journal articles.

Figure 25: Distributions of the within and between indicators for *The Library*, core books and journal articles respectively. Histograms are normalized, the lines are the kernel density estimations.

The correlations among indicators further elucidate this preliminary distinction, as shown in Table 14. For the first dataset, the within, topicality and bridging indicators act relatively in uniformity: they rise if the connectivity action of the core source is localized, they lower (and between indicator rises) if it is more global. In the second dataset this relation stands only for the within and topicality indicator, while bridging indicator loses correlation. For the last dataset, the relation lowers even more, at times becoming negative, and the main positive correlation is now between the within indicator and in-degree values. These results confirm the existence of two distinct ways of "being core": with a *global or localized connectivity action.* In the case of the first dataset, a local action can happen: within communities, and usually within a specific one, or bridging a given pair of communities. This was also clear from Figure 24a, as the between indicator is symmetric around zero to the within one, topicality is clearly positively correlated with the within indicator. A global action entails instead acting between several of these community pairs. For the last dataset, the action of the core (in this case, journal articles), is essentially only local either to a given community or to several communities, with few exceptions.

Table 14: Correlations of the indicators and the in-degree of the core sources. The between indicator is omitted as superfluous. Pearson: top-right. Spearman: bottom-left.

Dataset	Indicator	Within	Topicality	Bridging	Indegree
Books History Venice	**Within**	1	0.86	0.62	-0.13
	Topicality	0.92	1	0.38	-0.08
	Bridging	0.63	0.37	1	-0.19
	Indegree	-0.03	0.02	-0.14	1
The Library 1981-2016	**Within**	1	0.72	-0.01	0.05
	Topicality	0.7	1	0.08	0.06
	Bridging	0.01	0.09	1	0.11
	Indegree	0	0.2	0.09	1
All History 2005-2015	**Within**	1	0.44	-0.15	0.18
	Topicality	0.31	1	-0.01	0.44
	Bridging	-0.42	0.01	1	0.05
	Indegree	0.12	0.57	0.08	1

By taking a look at the top core sources for each dataset, according to every indicator, we can get a more concrete idea of the different structural roles core sources can have. Lists are provided in the Appendix: Tables 15, 16 and 17 for datasets one, two and three respectively. In general, the top of the between indicator indeed captures the sources most transversal to several scholarly communities. In the case of the first and second datasets, these are usually cornerstone monographs or reference works, in the last dataset we find instead methodological papers. Other indicators follow instead the behavior previously discussed. For example, the top sources by within and topicality for the first dataset on Venice, contain sources specific to the art and architecture history of the city.

Our results thus point to the presence of a core literature which is, where applicable, mostly composed of books. Book citations indeed seem to considerably rise the connectivity of the bibliographic coupling network. We also highlighted at least two structural actions that the core literature can play: a *local action*, connecting within one or more communities, and a *global action*, connecting across communities. The global one is mainly performed by some monographs and reference works, more rarely by journal articles. As it stands now, we are left wondering if the core literature is also dependent on the citing publications. Despite this remaining an open question, it would not appear so. Our second dataset, considering article citations to both WoS source and non-source items, effectively shows a mix behavior, in-between the first and third datasets.

5.4 Discussion

The accumulation of knowledge is a topic of great interest in bibliometrics. In the humanities, such is the variety of publication venues, typologies and languages, that it is difficult to disentangle the effects of each component of this *multidimensional system of knowledge*. The core literature is a particularly important element in this process. We defined it as the most cited sources of a given representative dataset, and asked the following questions: a) is there a well-defined core literature in history at different publication levels? b) What structural role does it play in the definition of the intellectual landscape of citing publications?

In order to answer such questions, we considered the bibliographic coupling network of citing publications, and their partitions into communities. Four indicators were introduced: a) the *within indicator* maps the action of a core source into connecting vertices within communities; b) the *between indicator* maps this action between pairs of communities; c) the *topicality indicator* assesses the proportion of the within action happening in a unique community; d) the *bridging indicator* assessed the proportion of the between action happening in a unique pair of communities. All indicators account for the skewed effect of the in-degree of the core literature by filtering it out using a null configuration model. The proposed method can be used in general to investigate the structural contribution of vertices after network projections such as bibliographic coupling. Three datasets were considered in order to explore the role of core literatures from different perspectives: a dataset of book to book citations, a dataset of journal articles to all source and non-source WoS items they cite, and a dataset including all journal article to journal article citations in history, as indexed in WoS over eleven years (2005-2015).

The main result is that the core literature, clearly emerging in all datasets under consideration, has at least two distinct structural effects on the intellectual landscape of the citing publications. This effect can either be *localized*, by rising the connectivity within one or several communities, or at times a specific pair of communities, or it can be *global*, by rising the overall connectivity of the landscape, across communities. We also found that a global action is usually performed by core sources which are well-known scholarly monographs, works of reference or primary sources. Local action can instead by performed by all kinds of core sources, and especially so by journal articles. The global action of core sources is reminiscent of the *strength of weak ties* in social networks [Granovetter, 1973], where a weak tie acts to connect between social groups thus providing an important, because otherwise weaker or absent link. As a result, the intellectual landscape of historians becomes better connected by considering citations to books, which at times span more broadly, and not just journal articles, which usually remain known within specific scholarly communities. Indeed, where applicable, the core literature is mostly composed of books. This result clarifies that both monographs, or works of interpretation, and reference works, or works on evidence, can play a crucial role with respect to the global intellectual organization of a research area.

There are indeed some limitations or left-open questions. First of all, the datasets which were used here present obvious limitations in that they are of sensibly different scale and coverage. Unfortunately, their use reflects the poor availability of citation data for the arts and humanities, and should compel further work in this direction to expand on these preliminary results. Another open question relates to the effective role of the core literature for a scholarly community, namely to what is the performative role of core sources for a given community. Is it more perfunctory, or still intellectually relevant? Taking a "Mertonian" approach and assuming most citations to the core literature to represent effective recognition of intellectual borrowing, the core literature might represent a form of loose or soft paradigm (Kuhn), keeping the community connected. Taking a constructivist approach, these citations could also be seen as perfunctory, as in signals of social membership within a group or otherwise persuasion, with little to no intellectual relevance [Bornmann and Daniel, 2008]. This question too, remains open and awaits further work.

Appendix

Table 15: Top-five core sources per indicator: Books History Venice.

Indicator	Value	Author	Year	Title
Within	0.49	Zanetti, A.M.	1733	Descrizione di tutte le pubbliche pitture della città di Venezia
	0.48	Temanza, T.	1778	Vite dei più celebri architetti e scultori veneziani
	0.47	Zanotto, F.	1856	Nuovissima guida di Venezia
	0.47	Aymard, M.	1966	Venise, Raguse et le commerce du blé
	0.46	Scamozzi, V.	1615	L'idea dell'architettura universale
Between	0.06	Filiasi, G.	1811	Memorie storiche de' Veneti
	0.06	Monticolo, G.	1896	I capitolari delle Arti veneziane
	0.06	Soranzo, G.	1895	Bibliografia veneziana
	0.06	Canal, M.	1972	Les estoires de Venise
	0.06	Molmenti, P.	1973	La storia di Venezia nella vita privata
Topicality	0.52	Scamozzi, V.	1615	L'idea dell'architettura universale
	0.52	Zanetti, A.M.	1771	Della pittura veneziana
	0.51	Ridolfi, C.	1914	Le maraviglie dell'arte
	0.51	Zanotto, F.	1856	Nuovissima guida di Venezia
	0.51	Temanza, T.	1778	Vite dei più celebri architetti e scultori veneziani
Bridging	0.58	Moschini, G.	1815	Guida per la città di Venezia
	0.55	Borsari, S.	1963	Il dominio veneziano a Creta
	0.54	Preto, P.	1975	Venezia e i Turchi
	0.54	Rösch, G.	1989	Der venezianische Adel bis zur Schliessung des Grossen Rats
	0.54	Pensolli, L.	1970	La gerarchia delle fonti di diritto nella legislazione medievale veneziana

Table 16: Top-five core sources per indicator: *The Library* 1981-2016.

Indicator	**Value**	**Author**	**Year**	**Title**
Within	0.72	Blagden, C.	1977	The Stationers' Company: A History, 1403-1959
	0.72	Plomer, H.R.	1925	Wynkyn de Worde & His Contemporaries
	0.72	Ker, N.R.	1987	Medieval libraries of Great Britain
	0.71	Shakespeare, W.	1606	King Lear
	0.70	Brusendorff, A.	1925	The Chaucer Tradition
Between	-0.04	Carter, H.	1975	The Oxford University Press
	-0.09	Venn, J.A.	1951	Alumni Cantabrigienses
	-0.1	McKerrow, R.B.	1927	An Introduction to Bibliography for Literary Students
	-0.11	Stow, J.	1908	A survey of London
	-0.15	McKenzie, D.F.	1978	Stationers' Company Apprentices
Topicality	0.58	Oldham, J.B.	1952	English Blind-Stamped Bindings
	0.58	Greg, W.W.	1967	A Companion to Arber
	0.57	Plomer, H.R.	1925	Wynkyn de Worde & His Contemporaries
	0.55	Hodnett, E.	1935	English Woodcuts 1480-1535
	0.54	Maxted, I.	1977	The London book trades, 1775-1800
Bridging	0.85	Smith, J.	1755	Printer's Grammar
	0.85	Arber, E.	1903+	The Term catalogs
	0.85	Morrison, P.G.	1955	Index of Printers, Publishers and Booksellers
	0.84	Foxon, D.F.	1975	English Verse, 1701-1750
	0.84	Lowry, M.	1979	The World of Aldus Manutius

Table 17: Top-five core sources per indicator: All History in WoS 2005-2015.

Indicator	Value	Author	Year	Title
Within	0.96	Karr, R.D.	1998	"Why should you be so furious?": The violence of the Pequot War
	0.96	Williams, S.	2005	Poor relief, labourers' households and living standards in rural England c. 1770-1834: a Bedfordshire case study
	0.96	Holquist, P.	2010	"In Accord with State Interests and the People's Wishes": The Technocratic Ideology of Imperial Russia's Resettlement Administration
	0.95	Krige, J.	2006	Atoms for peace, scientific internationalism, and scientific intelligence
	0.95	Runia, E.	2007	Burying the dead, creating the past
Between	-0.11	Aslanian, S.D.	2013	AHR Conversation How Size Matters: The Question of Scale in History
	-0.15	Harvey, D.	1990	Between Space and Time: Reflections on the Geographical Imagination
	-0.16	Stoler, A.L.	2006	On Degrees of Imperial Sovereignty
	-0.16	White, H.	1984	The Question of Narrative in Contemporary Historical Theory
	-0.18	Mann, G.	2005	Locating Colonial Histories: Between France and West Africa
Topicality	0.8	Bradley, J.	2002	Subjects into citizens: Societies, civil society, and autocracy in tsarist Russia
	0.8	Nora, P.	1989	Between memory and history, les lieux-de-memoire
	0.79	Weitz, E.D.	2008	From the Vienna to the Paris System: International Politics and the Entangled Histories of Human Rights, Forced Deportations, and Civilizing Missions
	0.79	Spear, T.	2003	Neo-traditionalism and the limits of invention in British colonial Africa
	0.79	Werner, M.	2006	Beyond comparison: Histoire croisee and the challenge of reflexivity
Bridging	0.99	Scott, J.W.	1991	The evidence of experience
	0.99	Foucault, M.	1986	Of other spaces
	0.99	Huntington, S.P.	1993	The clash of civilizations
	0.99	Spear, T.	2003	Neo-traditionalism and the limits of invention in British colonial Africa
	0.99	Greif, A.	1993	Contract enforceability and economic institutions in early trade: the Maghreb traders coalition

6 The changing intellectual organization of historians

Diachronic change is crucial when reasoning about the intellectual organization of a scholarly community, which is indeed an ever changing object. Philosophers and sociologists of science have amply reasoned about this aspect. Perhaps the most notorious theory of change in science is Kuhn's paradigm shift [Kuhn, 1996], where science is pictured as developing through a sequence of perception-altering revolutions instating new paradigms of scientific inquiry destined to last for longer periods of normal science. Some sociologists, including Collins [1975], Whitley [1984] and Fuchs [1992], have instead emphasized two other aspects influencing the way scientific communities change, among others: the degree of mutual dependence among scholars and the degree of task uncertainty [Chen and Song, 2017, Ch. 2]. These influence, in particular, the possible development of a community towards increased intellectual *specialization or fragmentation*. Adopting Fuchs' terminology in his theory of scientific change [Fuchs, 1993a] (also see Section 2.6.2), we consider a community to develop towards increased specialization if it possesses a high degree of mutual dependence among scholars, for example via a shared theoretical framework and set of questions or centralized research infrastructure, and a low degree of task uncertainty, so that most tasks are repetitive and their outcomes predictable. Conversely, a fragmented community possesses a low degree of mutual dependence and a high degree of task uncertainty, such is the case, according to Fuchs, for most of the social sciences and the humanities.

The problem of the effect of the accumulation of new literature in historiography is often discussed by historians as one of (perceived) *over-specialization*, thus with a negative connotation. In particular, during and after periods of sustained growth, scholars tend to lament the rise of specialization as in the narrowness of new publications, and the effects it has on the fragmentation of the field and on the scope of questions it tackles [Tyrrell, 2005; Colavizza, 2018]. More to the point, it is an open question how scholars, in particular in the humanities, react and adapt to the vertiginous amount of literature currently being produced [Bornmann and Mutz, 2015], and to its rising digital discoverability and availability [Evans, 2008; Larivière et al., 2009]. In this chapter we therefore consider two related questions: how can we measure the degree of specialization/fragmentation of a research area? And then, how are patterns of specialization/fragmentation changing over time more broadly in the discipline of history? We will assume Fuchs to be right in considering the humanities as fragmented, and history among them, leaving the question whether we might talk about specialization or fragmentation, or something else entirely, for the next chapter instead.

A preliminary design choice to make when studying the dynamics of a research area relates to the scale of analysis. By considering a single, possibly small community, it might be possible to analyze a longer span of time at a more granular level, e.g. using our Venetian dataset. Another option is a large-scale analysis relying on databases such as the Web of Science (WoS) or Scopus, which would guarantee to consider (many) more observations, but over a shorter span of time and with a focus on journal articles. We take an intermediate path here, by focusing on five areas of research or specialisms in history: economic history, social history, history of science, history of medicine and general English history. We do so by using data from WoS, representing each specialism using a set of three journals each, considered from the early 1950s to 2016 included, as permitted by data availability. Despite the focus on journal articles as citing publications, we consider citations to both source and non-source items (i.e. indexed in WoS or not). The focal point of attention will be the bibliographic coupling networks of each specialism so represented, considered over subsequent intervals of time. In particular, *fragmentation will be mapped to network connectivity*, in an attempt to detect whether it is rising, stable or declining over time. Three bibliographic coupling networks will be considered: i) of article citations (or reference overlap, the traditional one), ii) of author citations (identical but considering authors instead of articles as nodes), iii) of article textual similarity (measured over each article's title and abstract). We thus use bibliographic coupling in a generic way, to name networks where the nodes are scholarly publications, and the edges are determined by some similarity criteria among any two publications.

This chapter starts by introducing the methods used to construct the afore-mentioned bibliographic coupling networks, and those used to compare the networks over time, measuring network connectivity. These methods will also be used in Chapter 7, to cast a comparison over a variety of specialisms in the humanities and the sciences. Further, data and results are discussed, showing how the connectivity of these different specialisms in history is decreasing over time, pointing to a possible gradual but steady rise in the intellectual fragmentation of historiography as a whole.

6.1 Method

We start by defining the construction of the bibliographic coupling networks. Take $B = (V, E, w)$, the weighted bibliographic coupling network made of the publications of a given specialism, where $w : E \rightarrow \mathbb{R}^+$ is a function mapping each edge to a positive weight. W is the weighted symmetric adjacency matrix representing the graph, where $W_{i,j} = W_{j,i} = w(e_{i,j})$ if there exist an edge between vertices i and j, that is to say $e_{i,j} \in E$, 0 otherwise. Under this general setting, the edges and their weights can be established in a variety of ways. We consider three of them here: reference overlap of articles and authors (i.e. traditional bibliographic coupling as in Kessler [1963]) and textual similarity.

For reference overlap, we consider as the edge weight function w the cosine similarity calculated over the references that two publications i and j have in common:

$$w(e_{i,j}) = \frac{R_{i,j}}{\sqrt{R_i}\sqrt{R_j}} \tag{9}$$

Where $R_{i,j}$ is the number of references in common between i and j, R_i the number of references of i. We stress that we consider unique references, not their frequency (number of in-text references, or mentions). The cosine similarity is appropriate as it allows to evenly compare the weight of edges among publications with varied reference list lengths. Author to author bibliographic coupling networks are constructed by considering all (unique) references to publications made by an author within the given specialism and time period.

We base the textual similarity among two papers on the BM25 measure, widely adopted to rank documents for the purpose of information retrieval and document clustering [Sparck Jones et al., 2000a,b]. This measure has already been applied to assess the textual similarity of scientific publications (e.g. Boyack et al. [2011]; Colavizza et al. [2018a]), and it improves on simpler tf-idf by explicitly accounting for document lengths. Each publication text – in our case the concatenation of title and abstract – is reduced to lower case and split into tokens, further eliminating punctuation and then tokens of just one alphanumeric character. Given a publication i and another publication j, the BM25 similarity is calculated as:

$$s(i,j) = \sum_{z=1}^{n} IDF_z \frac{n_z(k_1+1)}{n_z + k_1\Big(1 - b + b\frac{|D|}{|\overline{D}|}\Big)}$$

where n denotes the number of unique tokens in i, n_z equals the frequency of token z in publication j, and $n_z = 0$ for tokens that are in i but not in j. k_1 and b have been set to the commonly used values of 2 and 0.75 respectively. $|D|$ denotes the length of publication j, in number of tokens. $|\overline{D}|$ denotes the average length of all publications in the dataset. The IDF value for every unique token z in the dataset is calculated as:

$$IDF_z = log\left(\frac{N - p_z + 0.5}{p_z + 0.5}\right)$$

where N denotes the total number of publications in the dataset and p_z denotes the number of publications containing token z. IDF scores strictly below zero are discarded to filter out very commonly occurring tokens. BM25 is not a symmetric measure. We thus obtain a symmetric measure for the similarity of publications i and j, the value is the weight of the edge connecting them in B, as follows:

$$w(e_{i,j}) = \frac{s(i,j) + s(j,i)}{2} \tag{10}$$

While the BM25 textual similarity is calculated for every publication pair, the IDF scores and $|\overline{D}|$, the average length of all publications, are calculated and shared globally over all datasets. We refrain from further normalizing the similarity scores, in order to allow for comparisons across datasets.

Connectivity and giant component A connected component of B is a sub-graph whose nodes are all connected, i.e. there exists a path between every pair of nodes in the component. An isolated node is a node that is not connected to any other node (hence representing a singleton connected component). The giant component is the largest connected component in the number of nodes it contains [Newman, 2010, 142-3].

In order to explore the connectivity of the bibliographic coupling networks introduced above, we measure the proportion of connected components over the total possible (Eq. 11), at steps in which we remove all edges below a certain weight threshold. This method allows to assess the strength of edge weights in the network, and the behavior of the connected components as the network becomes increasingly disconnected. This procedure can be considered as an analysis of a form of t-edge-connectivity, where a component is considered as connected only if it is a connected component by considering edges of weight at least equal to t. Alternatively, it is a form of bond percolation where edges are removed deterministically according to their weight.

Given an edge weight threshold t, we are thus interested in a measure which is calculated at increasing t over networks B:

$$c(t) = \frac{C^t}{N} \tag{11}$$

Where N denotes the number of publications, equivalent to the number of nodes in the network and also equal to the number of connected components in the disconnected network; C^t denotes the number of connected components after removal of edges with weight below t. It is worth pointing out that the measure consider the structure of the network after the removal of some edges, but does not account for the weight of the remaining edges. This might appear as a limitation, but in practice the analysis of the whole process accounts for both structure and relative weight of edges at every step. Alternative measures, such as algebraic connectivity or k-connectivity [Newman, 2010], did not yield complementary results of note.

6.2 Dataset

Five specialisms in history are considered, covering mainstream areas of research. For each, three journals are chosen to represent trends in the research being published therein. Due to the limitations of WoS and in order to maximize coverage over time, the journals were selected to be internationally renown ones in English. An overview of the dataset is given in Table 18. Limitations in the coverage of WoS (which starts at best in 1956), or the date of initial print of some journals, determines an uneven data availability for the initial years under consideration. In any case by the 1980s all journals are active and indexed. With respect of the presence of abstracts, the situation is less fortunate: most journals either do not have them, have them unevenly or they are not present in the database. We therefore consider only one journal per specialism with respect to textual similarity, the one with better abstract availability (marked by an asterisk in Table 18).

The coverage of references allows to consider six time periods for both article and author bibliographic coupling networks: up until 1969 included, 1970-1979, 1980-1989, 1990-1999, 2000-2009, 2010-2016. With respect to textual similarity, only three periods are considered instead: 1999-2004, 2005-2010, 2011-2016.

The data was downloaded directly from the WoS interface and processed in order disambiguate all references and authors. Reference lists were first parsed to extract and clean every reference as follows. First, every reference was split into author, publication year and the rest of its text (mainly, the title). The title was then trimmed from any page, number, issue or volume information. Lastly, all anonymous references (in the author field) or references without a publication year were discarded. The second step entailed the creation of a single global reference dictionary for the whole dataset. In order to do this, every reference was compared with every other reference, and a match was established if all the following conditions were verified: a) the first three characters of the author and title fields matched exactly (to lower case); b) the Jaro-Winkler similarity between author fields was equal or greater than 0.9; c) the similarity of the title fields was equal or greater than 0.85 and the publication year matched exactly, or the similarity of the title fields was equal or greater than 0.95 (useful in the case of different editions of the same work). A total of 777'894 references were considered, resulting in 443'561 disambiguated references.

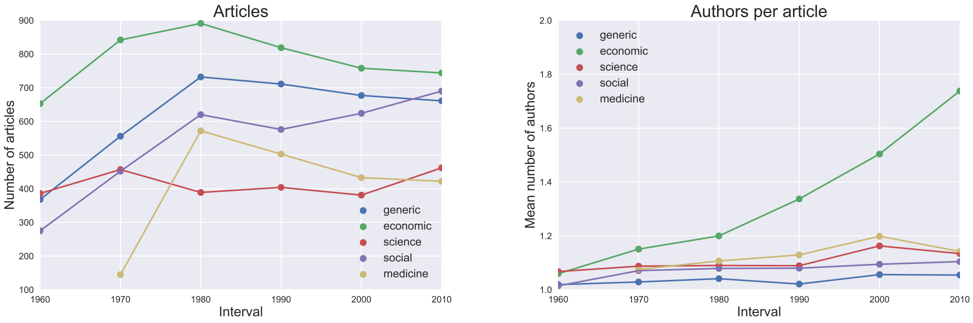
Article authors were instead disambiguated field by field. A match was established if the (lower cased and punctuation stripped) surnames had a similarity strictly higher than 0.95 and names strictly higher than 0.9, using the Jaro-Winkler measure. The number of unique authors per field is given in Table 18, highlighting how the communities of historians of science and of medicine are smaller than the rest, according to this dataset.

An important element to consider in order to compare the connectivity of networks over time is their relative similar size. The number of articles per specialism is given in Figure 26a, showing how from the second period, the number of articles is relatively stable within each specialism (with the minor exception of the second period of the history of medicine). A similar comparability has to hold for the number of authors and especially for the relative importance of co-authorships. Co-authorship often determines edges of weight one in the author bibliographic coupling network, therefore it has

Table 18: Summary of datasets. Journals marked with an asterisk are used for creating the text similarity networks, due to their better availability of abstracts.

Field	**# authors**	**Journal**	**Coverage**	**# articles**	**# articles with abstract**	**Abstracts since**
General English History	2935	English Historical Review	1956-2016	1220	-	3
		The Historical Journal*	1966-2016	1784	1999	606
		Journal of British Studies	1975-2016	701	2013	117
Social History	2908	Past&Present	1956-2016	1411	-	2
		Journal of Social History*	1967-2016	1344	1991	737
		Social History	1972-2016	481	1999	138
Economic History	3399	Explorations in Economic History	1969-2016	1065	1994	544
		Journal of Economic History*	1956-2016	1996	1991	734
		Economic History Review	1956-2016	1646	1992	639
History of Science	1872	Isis*	1956-2016	1277	1997	406
		History of Science	1987-2016	377	2014	58
		Annals of Science	1966-2016	825	1991	337
History of Medicine	1727	Bulletin of the History of Medicine*	1977-2016	738	2001	262
		Journal of the History of Medicine and Allied Sciences	1978-2016	550	2003	181
		Medical History	1978-2016	787	2011	136

a strong influence on connectivity. As shown in Figure 26b, co-authorships are very rare and stable for all specialisms, with the important exception of economic history, where they significantly rise over time. Interestingly, economic history is strongly related to economics and the social sciences, and might be borrowing some traits of those communities [Henriksen, 2016]. Connectivity results for author networks will therefore be given with and without economic history, to account for its different behavior in this respect.



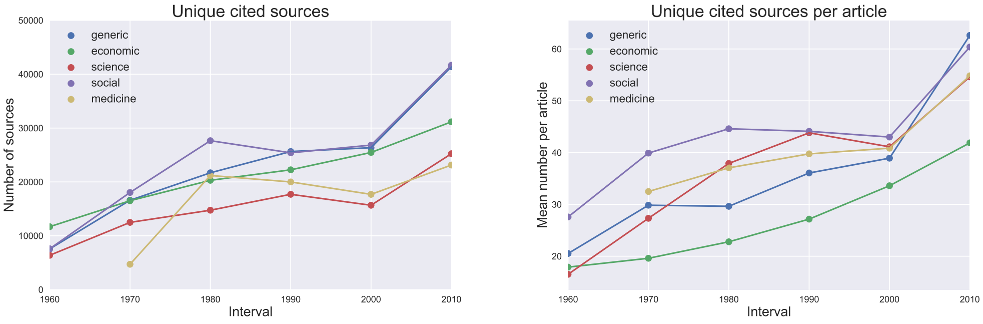
(a) Number of articles.

(b) Number of authors per article.

Figure 26: The number of articles and number of authors per article, over specialisms and periods. Every period is individuated by its start date (the first period in fact starts in 1956). The datapoints are plotted as circles, while the lines uniting them are just meant to aid the reader.

A second aspect to consider is the global number of unique references and especially the mean number of unique references per article, both given in

Figure 27. They show a clear trend towards a progressive increase, especially pronounced over the last period (2010-2016), shared with some differences among all specialisms. If the length of reference lists rises over time, this can have an influence over the distribution of the cosine similarities among articles, as per Eq. 9. In particular, if the raw number of shared references remains constant but the length of reference lists rises, the cosine similarity will become smaller. Therefore there are two ways for the cosine similarity to lower: when there are fewer shared references over comparable reference list lengths, and when the number of shared references remains stable or anyway rises less rapidly than the length of reference lists.



(a) Unique cited sources per specialism. (b) Unique cited sources per article.

Figure 27: The number of unique cited sources per specialism and per article. Every period is individuated by its start date (the first period in fact starts in 1956). The datapoints are plotted as circles, while the lines uniting them are just meant to aid the reader.

A last element worth considering is the Price index [De Solla Price, 1970], given in Figure 28. The Price index here is the proportion of cited sources published maximum within 10 years from the cited one, and conveys an idea of the age of the cited literature of a specialism. Two specialisms, social and economic history, starting from a relatively high Price index in the 1950s to 70s, have been rapidly falling as their literature grew old and numerous. Conversely, the other three specialisms have seen a rise of relative stability of their Price index, which was sensibly lower to begin with. The index over the last period is broadly comparable to the one calculated for the Venetian dataset of book to book citations, discussed in Chapter 4, highlighting once more the importance of old literature in history.

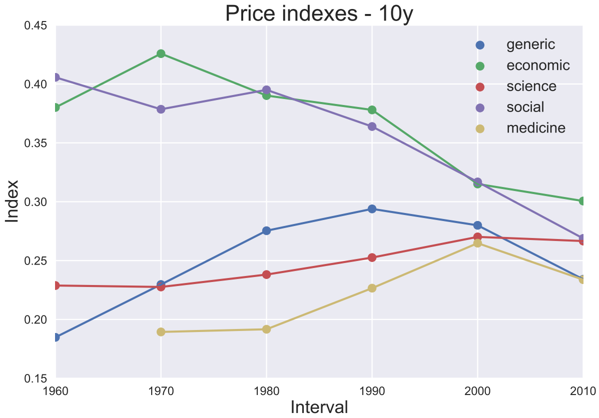


Figure 28: The Price index, over specialisms and periods. Every period is individuated by its start date (the first period in fact starts in 1956). The datapoints are plotted as circles, while the lines uniting them are just meant to aid the reader.

6.3 Results

We are interested to see whether the general connectivity of the bibliographic coupling networks of different specialisms in history is changing, and especially if it is declining, over time. Let us thus consider Eq. 11, averaged over specialisms and plotted at increasing thresholds over periods of time. As a reminder, by increasingly removing edges whose weight is below the given threshold, the network collapses into many small connected components. We consider the speed of collapse in order to see whether the connectivity of these networks is declining over time.

Starting with the classic reference overlap bibliographic network (Eq. 9), results are shown in Figure 29. It can be clearly seen that progressively, from the 1980s at least, the mean and median connectivity is declining. Intuitively, from the first period to the last, the average bibliographic coupling network considered at threshold 0.1 collapses from 40% to 80% of the total possible connected components. These results are also remarkable for showing a gradual decline in connectivity, as if this process were not driven by conjunctures or specific events, but from a steady change, suggesting that at the article level, the decline in connectivity might be due both to a decline in the raw number of shared references and a rise in the number of non-shared ones.

Results at the individual specialism level, which are here omitted for brevity, highlight that general, social and science history show more marked decline in connectivity, whilst economic history and the history of medicine are somewhat more resilient to it, albeit participating in the same general trend.

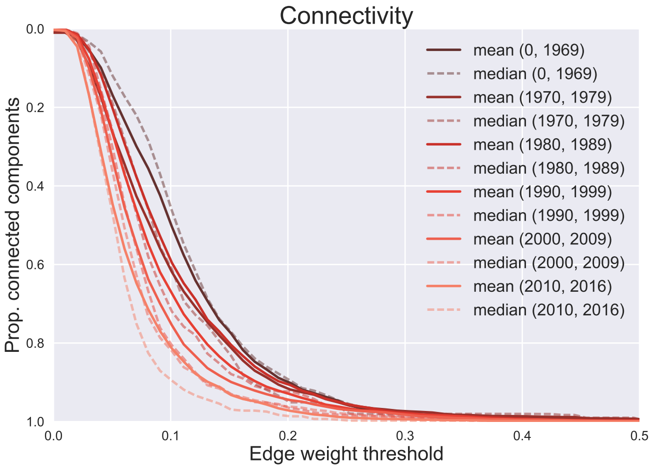
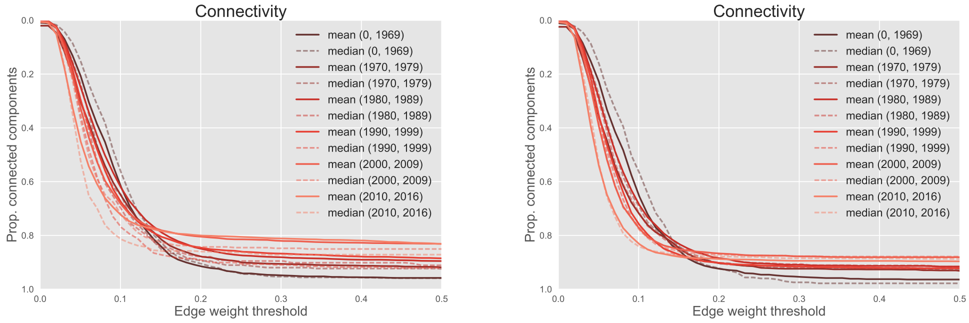


Figure 29: The connectivity of reference overlap bibliographic coupling networks over time. Please note the y axis is reversed and goes from 1 to 0 instead.

We then consider author bibliographic coupling networks, where authors are the nodes and connections are established if they share references among themselves. We need to account for results with and without economic history separately, due to the rising frequency of co-authorships in this domain, as previously discussed. Results are given in Figure 30. We can see that results follow alongside reference overlap networks, albeit with less marked effects. The decline in connectivity is particularly sensible over the last period, where the number of unique cited sources has been sensibly rising (cf. Figure 27). At the author level therefore, we might be witnessing a stability in the raw number of shared references, but a rise in the number of non-shared ones too, over recent times. Essentially, authors seem to refer to more, non-shared sources. Finally, we can appreciate how important the impact of co-authorships is on the general connectivity of an area of research.

Economic history stands out in this respect, but the general effect is still present when discarding its signal, as per Figure 30b.

Results at the specialism level show how economic history and the history of medicine have stable or even slightly rising connectivity at the author level, contrary to the rest of the specialisms under consideration.



(a) With all specialisms.

(b) Without economic history.

Figure 30: The connectivity of author reference overlap bibliographic coupling networks over time. Please note the y axis is reversed and goes from 1 to 0 instead.

Lastly, textual similarity bibliographic coupling networks are considered in Figure 31. We remind that these results only consider the title and abstract as text representation for an article, using data for only one journal per specialism (cf. Table 18) and considering fewer, more recent periods of time. Interestingly, results point to a stable or rising connectivity at this level. All journals participate in this trend, with Isis showing a particularly strong rise in similarity from the first to the second period.

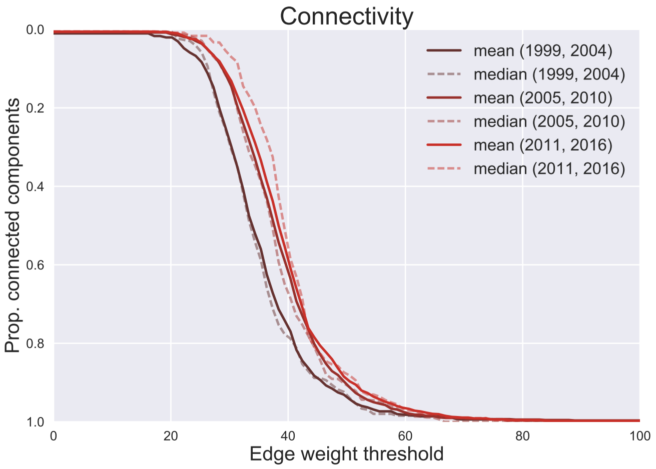


Figure 31: The connectivity of textual similarity bibliographic coupling networks over time. Please note the y axis is reversed and goes from 1 to 0 instead.

In summary, we have seen that the general connectivity of article reference overlap bibliographic coupling networks is gradually and steadily declining over time. The author reference overlap connectivity is also in general declining, but only slightly, with some exceptions and an important counter effect provided by the rise in the number of co-authorships, particularly marked for economic history. Lastly, the textual similarity connectivity is stable.

6.4 Discussion

To recapitulate, this chapter explored the extent in which the intellectual cohesion of different specialisms in history is changing over time. The proposed method was that of exploring the connectivity of bibliographic networks of articles and authors using reference overlap and textual similarity. Our results highlight some aspects of importance.

The declining connectivity of article bibliographic coupling networks can be due either to a decline in the raw number of shared sources or to the rising number of non-shared ones, possibly both. The connectivity of author bibliographic coupling networks is less affected by this trend, in part due to the counter effect of co-authorships, especially strong in economic history. It is also evident that author networks decline in connectivity more markedly over the last period, when a sensible rise in the number of unique references is recorded (i.e. reference lists get longer). In accordance to what has been found for the sciences and social sciences, the concentration of citations might be lowering in recent times in part due to the rising amount of references made per article. This phenomenon might explain the reported trend, albeit previous evidence points to a general stability in the concentration of citations for the humanities [Larivière et al., 2009]. Lastly, the persisting levels of connectivity in the textual similarity networks might highlight a steady effort in integrating research results within shared narratives.

Previous studies have shown how historians are not narrowing the scope of their attention with respect to primary sources, but are indeed broadening their topical and methodological perspectives and interests [Colavizza, 2018]. The results of this chapter can therefore be tied together into a possible interpretation: the undeniable emergence of different directions of research in historiography over recent decades, coupled with a growing amount of published results, is determining a gradual decline of the citation connectivity of publication networks as scholars produce more fragmented or specialized research. Nevertheless, at the level of scholars this decline is only marginal, if happening at all, thanks in part to a generally modest but in some specialisms more sensible rise in collaborations as evidenced by co-authored results. Another way for historians to tie their results together is their reliance on a shared vocabulary and set of narratives.

History thus emerges as a discipline where the pressures of growth are determining i) a decreasing reference connectivity of publications within specialisms and, to a lesser degree, of scholars, ii) a perhaps only too tenuously growing propensity for collaborations, iii) a constant reliance on shared narratives in order to integrate new results together. These results highlight an emergent property of the intellectual organization of history, faced with the gradual growth of accumulated knowledge. According to Jones [2009], the accumulation of knowledge leads to an increasing educational burden, a narrowing of expertise (rise in specialization) and an increased propensity for teamwork within a cohesive social system. Importantly, these general strategies can differ in magnitude cross-sectionally, likely in a way which is correlated to the degree of consensus within the specialism or even the field, as we briefly discussed in the introduction. Considering history as a discipline, this leads us back to our distinction between interpretive work and work on primary evidence (such as catalogs, critical editions or databases) [Ziman, 1968]. Whilst the latter might lend itself more easily to consensus on methods and results, thus possibly going further into exploring specialization and teamwork within a shared methodological framework, the former is likely to put a premium on a variety of ways to interpret novelty and originality [Guetzkow et al., 2004]. This attitude to the reception of interpretive work is also evidenced by the tendency of humanities authors to take explicit points of view and emphasizing their own contributions [Hyland, 2006]. As a result, "citation fragmentation" of interpretive works (such are most journal articles) might indeed occur, but not resulting from specialization nor paraleled by a rising propensity for teamwork. Intellectual fragmentation might slow scientific progress by limiting a scholars' influence among their peers [Balietti et al., 2015], and lead to the outcome that the integration of new results into the body of disciplinary knowledge becomes increasingly reliant on the principal form of expression in historiography: narratives.

To further elucidate this point, in the final chapter we will cast a comparison between history and other disciplines in the humanities and the sciences, in order to clarify how and to what extent they differ with respect to their intellectual organization.

7 A comparison of the sciences and the humanities

A long tradition of sociological research aims to understand the differences in the organizational and cognitive structure of scientific fields [Merton, 1974; Collins, 1975; Whitley, 1984; Becher, 1989; Fuchs, 1992]. This sociological tradition was in its earlier years intimately connected with the emerging field of bibliometric methods and applications, originated in the 1960s with the work of Storer and Price [Storer, 1967; De Solla Price, 1970]. For example, the Price index has played an important role in the early sociology of science [Zuckern and Merton, 1973; Cole, 1983] and is of continuous importance in scientometrics to this day [Wouters and Leydesdorff, 1994; Larivière et al., 2008]. However, the sociology of science and scientometrics have since the early 1980s largely drifted apart and attempts to reconcile them, or to reconcile the more theoretically inclined field of science and technology studies with scientometrics, have not had the desired effect [Leydesdorff, 1989; Luukkonen, 1997]. Recently, scholars have again argued for the need for interdisciplinary work bridging the sociology of science [Gläser and Laudel, 2016] or science and technology studies [Wyatt et al., 2016] with scientometrics. Within scientometrics, theoretical scholarship has been aimed at citation theories [Bornmann and Daniel, 2008; Tahamtan and Bornmann, 2018], which are crucial to understand differences in referencing behavior for the legitimacy of citation measures in evaluation procedures and practices [Wouters, 1999].

In this chapter, we take up these calls directly and explore ways to bridge the sociology of science with scientometrics, using science mapping methods to operationalise a specific sociological theoretical framework. The field of science mapping has developed network methods to analyze the cognitive structure of different fields, as well as their relative interdependence (see Section 2.5). Such network methods could be used to test sociologically informed hypotheses regarding the cognitive structure of scientific fields and differences between them [Fanelli and Glänzel, 2013]. One of the most influential sociological frameworks is presented in the work of Becher [1989] who developed a conceptualization of academic territories and their tribes, initially on his own and later with Paul Trowler [Becher and Trowler, 2001] (also see Section 2.6.3). Becher and Trowler argue that epistemic structures have both a cognitive and a social dimension and that communication practices of their tribes mirror (and thus reproduce) these structures. The framework has been critiqued extensively especially for its essentialist view, also by Trowler [2013, 2014], in particular for its assumption that knowledge claims are in essence hard or soft, pure and applied, which is arguably a sharp distinction. However, fully accepting this criticism, their work employs inventive distinctions, such as those of convergent versus divergent fields and of rural versus urban territories. The latter is the focus point of our analysis, as it proposes to jointly explain the social and intellectual organization of different research specialisms mainly via the axes of research topics and the amount of researchers each topic gathers. Essentially, rural areas organize in many, small topics, thus resulting in fragmentation, urban specialisms instead organize into few, more populated topics. This idea has some traction in scientometrics when distinguishing between communication patterns in the mathematical and natural sciences, and the arts and humanities [Hammarfelt, 2011, 2016]. Fully acknowledging that a spectrum of specialisms exists between these two conceptual opposites, this conceptualization allows to identify a broader and related set of characteristics that rural and urban specialisms might possess. Communication patterns, which include but are not limited to publication practices, accordingly give insight into these structures, and consequently into the differences between disciplines. Based on [Becher and Trowler, 2001, Ch. 6] we thus develop a number of hypotheses of what we can expect to observe in rural or urban specialisms.

In this chapter we therefore zoom out of history maximally, and consider it as a discipline part of the humanities, in comparison with other disciplines in both the sciences and the humanities. A comparative effort is important in order to put our previous results into perspective. At the same time, we take a theory-driven approach by explicitly operationalizing a conceptualization from the sociology of science in order to guide the empirical work. This chapter thus is organized as follows: we first discuss the operationalization, or reduce the conceptualization from Becher and Trowler [2001] to a set of observable objects and testable hypotheses. We then define how to test each hypothesis and discuss dataset and results. The main outcomes of this chapter is the clear distinction emerging between humanities and sciences with respect to both reference and textual connectivity, suggesting that their intellectual organization indeed differs substantially.

This chapter borrows from Colavizza et al. [2018b].

7.1 Operationalisation

The main distinction between *rural and urban specialism* is made based on the number of topics studied within a community at a given time – low for urban specialisms, high for rural specialism – and the "people-to-problem" ratio, meaning the number of researchers involved in a research topic at any one time – high for urban specialism and low for rural specialisms. We hypothesize:

1. Hypothesis 1: The number of topics being researched is high for rural specialisms and low for urban specialisms. In rural specialisms more, smaller topics are expected to be found, everything else being equal, while in urban specialism fewer, larger topics are expected to be found.

2. Hypothesis 2: Rural specialisms have a low people-to-problem ratio and urban specialisms a high people-to-problem ratio.

The authors subsequently suggest that this difference has implications for publication practices. They argue that rural authors have a broader scope intellectually and move more freely between topics. As there is no clear agreement about the core problems, in each publication the argument has to be embedded explicitly in the previous literature of the specialism, across topical boundaries. Therefore, these publications are on average longer, have more references and references are more evenly distributed across the specialism, as they are less focused on the specific topic. In urban specialisms, on the other hand, publications are shorter, contain less references and references are highly specialized, as there is no need to legitimize or contextualize the publication by referencing outside of the topic. We hypothesize:

3. Hypothesis 3: publications in rural specialisms are longer.

4. Hypothesis 4: publications is rural specialisms contain more references, both in absolute sense and relative to the publication length.

5. Hypothesis 5: references in publications from rural specialisms cover a larger variety of sources (e.g. more different journals).

6. Hypothesis 6: in rural specialisms there are comparatively more core publications that are shared beyond topics, making the specialism more reliant on them overall. This is less the case in urban specialisms,

where core publications are mostly restricted to a topic. By core we mean highly cited publications. Intuitively, there are more weak ties across topics in rural specialisms than urban ones, due to the need to embed arguments within the broader specialism and not just within the specific topic.

The rural and urban distinction also has implications for productivity and collaboration practices, although these are, according to Becher and Trowler [2001], primarily effects of higher competition in urban specialisms rather than resulting from its internal cognitive and social structure. They further argue that in a more competitive specialism, productivity is higher. Moreover, because there are many people working on similar problems, there is a heavy competition to be the first to solve research problems. In such competitive environments, there is a self-reinforcing incentive to work together, therefore the average number of authors is higher. We hypothesize:

7. Hypothesis 7: authors in rural specialisms publish less, but across a wider range of topics. Scholars in urban specialisms publish more but within a smaller range of topics.

8. Hypothesis 8: the average number of authors is higher in an urban specialism than in rural specialism, and there are more collaborations in urban specialisms.

The question why a specialism is urban or rural in the first place is a crucial one that we cannot answer in the present analysis. The authors themselves suggest that the amount of competition is the main reason for a specialism to become urban, and this can be the result of an emergent new research paradigm following a Kuhnian revolution, or be influenced by science policy [Becher and Trowler, 2001, 105-6]. There are a few more aspects in the urban and rural analogy which we cannot consider in this paper: the authors suggest that urban specialisms show a stiffer competition for resources (e.g. budget allocation, students, etc.), and have more rapid and heavily used (in)formal information networks.

The first step into the operationalization of the rural and urban conceptualization of the structure of scientific fields is to proxy its two basic units of analysis: *specialisms and topics.* A specialism is a group of people (a community) focusing on related topics of research which communicates this research internally through specialized journals, conferences and seminars.

A topic of research is a well-identified set of problems and related questions, recognized by the community as being of interest and part of it. For example, in the specialism of natural language processing, speech recognition is a topic. Topics can be individuated at different granularities.

> We proxy a *specialism* by considering a community producing publications which are a-priori well-individuated (e.g. by publication venue). We then proxy a *topic* as a well-connected cluster in the bibliographic coupling network of the publications published by authors active in the specialism.

Bibliographic coupling networks can be constructed in several ways, for example considering reference overlap or textual similarity between publications, as proxies for their relatedness. We consider a well-connected cluster to be a connected component with a minimum edge weight on every internal edge. We thus use connected components to approximate topics. A connected component is a sub-graph where every node is connected to other nodes by at least a path. In summary: we proxy a specialism by considering a set of externally grouped publications (for example all publications from a journal), a topic is then a specific connected component of such publications in the resulting bibliographic coupling network, with a minimum edge weight on every component edge. In so doing, we only aim at approximating topics. As we did in Chapter 6, we use bibliographic coupling in a generic way to name networks where the nodes are scholarly publications, and the edges are determined by some similarity criteria among any two publications. To be sure, other approaches might be considered, for example individuating topics using community detection methods or topic modeling on full texts. The main issue with these methods is that it is difficult, and ultimately involves judging whether topics are indeed coherent, to arrive at comparable topics across different specialisms. The proposed method preserves the benefit of simplicity of interpretation and does not require us to judge whether a topic should be identified as such but rather assumes that overlap in references identifies similarity between publications. In what follows we focus on networks of publications and require that specialisms possess a comparable number of publications each. An alternative would have been to consider networks of authors and require specialisms to be comparable in the number of active authors. The main reason we did not pursue this direction is the added complexity in accounting for the impact of co-authorship on bibliographic coupling networks of authors.

In this study, we focus directly on hypotheses 1 and 6, and on hypothesis 2 by implication, which we consider central to the theory and little explored in the literature. We consider hypotheses 3, 4 and 8 only using metadata, since these aspects have already been considered and largely confirmed in the literature. We do not consider hypotheses 5 and 7, since they would require different design or data. In particular, hypothesis 5 would require to first assess the variety of venues and publication typologies, likely higher in rural specialisms than in comparably-sized urban ones. Given our operative definition of specialisms and topics, we propose to operationalize the main hypotheses derived from the rural and urban analogy as follows:

- Number and size of topics (hp. 1): we remove edges at increasing weight thresholds. The connectivity of the network in terms of the number and size of its connected components gives us a way to measure the relative number and size of topics. According to hypothesis 1, rural specialisms will fragment into more topics given the same weight threshold than urban specialisms, as illustrated in Figure 32.

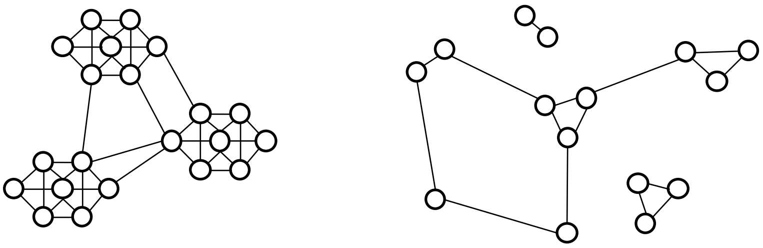


Figure 32: Illustration of the hypothetical topic granularity in rural and urban specialisms. Urban specialisms organize tightly in fewer, larger clusters (left); rural specialisms in more, smaller clusters (right). By removing low-weight edges (represented here by longer edges), rural specialisms will disconnect into more, smaller connected components.

- People-to-problem ratio (hp. 2): we operationalize the second hypothesis through the number of authors (people) active per topic (connected component) at the same edge weight threshold.

- Length of publications (hp. 3): we consider for each specialism the average length of publications in number of pages.

- Number of references (hp. 4): we consider for each specialism the average number of references, and the average number of references per page.

- Core publications (hp. 6): we compare the concentration of citations across the whole specialism to identify core sources that are highly cited. We measure the effect of core sources on the overall network connectivity by removing them in order of received citations or at random. We expect the global (out of topic) reliance on core sources to be greater in rural specialisms, thus the impact of removing them first to be comparatively lower on the overall network structure and size of the giant component at the specialism level. By impact here we mean the relative importance of sources into connecting the network, therefore a rural specialism will initially disconnect less rapidly by removing core sources first. Urban specialisms should instead be reliant on core sources at the level of large topics, less so globally.

- Collaboration (co-authorship, teamwork) and population (number of authors, hp. 8): both low for rural, high for urban specialisms. The best-known method to proxy collaborations are co-authorships. It has been already confirmed, and our study will too, that co-authorships are rarer and involving fewer authors in allegedly rural specialisms such as many in the humanities [Tsai et al., 2016].

7.2 Dataset

We selected ten specialisms, and corresponding datasets, within five disciplines (two specialisms each): history, computer science, astrophysics, literature and biology. Each dataset is extracted from Scopus and contains representative publications for every specialism over several contiguous years. These datasets are not comprehensive, something extremely difficult to achieve in general, but hopefully representative of the research published in the respective specialism. We used Scopus due to its better coverage of computer science conferences with respect to the Web of Science. The selection of journals has been reviewed by at least one domain expert. International and renown venues have been preferred, as follows:

1. Specialism *A1, economic history.* Research articles from the following journals: Explorations in Economic History, Journal of Economic

History, Cliometrica, Economic History Review, Business History.

2. Specialism *A2, history of science.* Research articles from the following journals: Social Studies of Science, Isis, Studies in History and Philosophy of Science, Studies in History and Philosophy of Modern Physics, Historical Studies in the Natural Sciences, Archive for History and Exact Sciences, History of Science, Annals of Science, History and Philosophy of Life Sciences, Technology and Culture.

3. Specialism *B1, computer science, neural networks and machine learning.* Conference papers from the annual conference on Neural Information Processing Systems (NIPS).

4. Specialism *B2, computer science, natural language processing.* Conference papers from the annual conference of the Association for Computational Linguistics (ACL).

5. Specialism *C1, astrophysics, solar system.* Research articles from the journal Icarus.

6. Specialism *C2, astrophysics, cosmology and astroparticle physics.* Research articles from the Journal of Cosmology and Astroparticle Physics (JCAP).

7. Specialism *D1, literature, classics.* Research articles from the following journals: Classical Quarterly, Mnemosyne, Hermes, Zeitschrift für Papyrologie und Epigraphik, Rheinisches Museum für Philologie, International Journal of the Classical Tradition, American Journal of Philology, Classical Journal, Classical Philology, Classical Receptions Journal, Quaderni Urbinati di cultura classica, Cambridge Classical Journal, Journal of Hellenic Studies.

8. Specialism *D2, English literature.* Research articles from the following journals: English Studies, Victorian Studies, Victorian Literature and Culture, Studies in English Literature, Review of English Studies, Studies in Philology, European Journal of English Studies, English Literary Renaissance, Studies in Romanticism, Journal of English Studies, International Journal of English Studies.

9. Specialism *E1, biology, neuroscience.* Research articles from the journal Neuron.

10. Specialism *E2, molecular biology.* Research articles from the journal Molecular Biology and Evolution (MBE).

The reason to select multiple journals for specialism A1, A2, D1 and D2 relates to their more diffuse publication practice (lower number of articles, higher number of book reviews per issue). As a consequence, a higher number of journals had to be selected in order to gather an overall comparable number of articles per year.

Summary statistics for the datasets under consideration are given in Table 19. The overall number of articles is comparable, yet other sensible differences emerge. In particular, articles are longer in history and literature specialisms, and they also possess fewer authors (with the partial exception of economic history), in agreement with hypotheses 3 and 8. The number of references varies greatly too, with computer science having fewer of them due to the shorter format of conference proceedings, but no clear trend is visible, contrary to what expected from hypothesis 4. In particular, there does not seem to be a clear distinction with respect to references per page, with literature possessing the lowest and biology the highest amount.

Table 19: Summary statistics for the datasets under consideration.

Statistic/Specialism	**A1-ec_hist**	**A2-hist_sci**	**B1-NIPS**	**B2-ACL**	**C1-icarus**	**C2-JCAP**	**D1-classics**	**D2-eng_lit**	**E1_neuron**	**E2_MBE**
Number of articles	1115	2379	2137	1904	2932	3760	1806	1159	2115	1827
Number of references	68’016	141’411	51’123	51’635	153’693	225’084	74’016	53’218	119’212	110’396
M(M) references per article	61.2(55)	62.6(52)	23.9(24)	27.1(26)	52.4(46)	59.9(53)	42.8(30.5)	46.2(42)	56.7(57)	60.5(58)
M(M) authors per article	1.8(2)	1.2(1)	3.1(3)	3.1(3)	5.2(4)	5.4(3)	1.1(1)	1.1(1)	7.5(6)	5.5(4)
M(M) pages per article	22.1(22)	16.3(11)	8.3(8)	7.5(9)	12.1(11)	17.9(17)	14.6(11)	19(19)	11.8(12)	10.5(11)
Number of articles 2016	147	350	-	232	416	527	263	150	342	247
Number of articles 2015	143	367	403	316	454	648	375	152	318	261
Number of articles 2014	137	357	411	286	431	658	229	160	343	277
Number of articles 2013	175	384	360	328	388	570	211	233	309	234
Number of articles 2012	178	352	370	188	402	538	308	207	286	261
Number of articles 2011	183	293	301	292	413	420	246	172	266	293
Number of articles 2010	152	276	292	262	428	399	174	85	251	254

Despite the fact that the size of the selected specialisms can be considered to be comparable in terms of the number of publications, it is not so with respect to the number of authors. Table 20 reports the number of author mentions and the number of unique authors for every specialism. The number of unique authors was calculated by merging authors with the exact same surname and forename initials, assuming homonymity to be a relatively marginal event within each specialism. Clearly the number of authors publishing in each specialism varies greatly, with literature and history numerically at the bottom end, biology at the higher end, confirming hypothesis 8 in this respect.

Table 20: Number of author mentions and unique authors per specialism.

Statistic/Specialism	**A1-ec_hist**	**A2-hist_sci**	**B1-NIPS**	**B2-ACL**	**C1-icarus**	**C2-JCAP**	**D1-classics**	**D2-eng_lit**	**E1_neuron**	**E2_MBE**
Number of author mentions	1442	2039	5685	4453	11'335	14'541	1437	999	11'321	7254
Number of unique authors	1102	1662	3134	2290	4808	5487	1180	936	9058	5552

Data acquisition From the Scopus interface, all relevant research articles or conference papers were downloaded, including their references. In order to include source and non-source items in our analysis merging references to the same object was need. To do so we proceeded as follows. Firstly, the authors were separated from the rest of the reference. Secondly, references without author were discarded. For two references to be merged into the same object cluster, three things need to happen: 1) the surnames of the first authors need to match; 2) the two lists of authors need to have a Jaro-Winkler score of 0.9 or above; 3) the rest of the reference text needs to have a Jaro-Winkler score above a threshold determined for each specialism/dataset. This last threshold is established empirically by finding the score yielding an accuracy of less than 0.5 in the 100 pairs of references to be merged with a score just below that threshold. Similarly, the 100 pairs immediately above the threshold must yield an accuracy above 0.5.[22] The intuition is that the accuracy of matches above the threshold rapidly improves, as it rapidly deteriorates below the threshold, therefore yielding acceptable results. The thresholds and accuracy scores for every dataset are reported in Table 21. The Jaro-Winkler measure is specifically designed to match short texts that might be misspelled at their end instead of at their beginning, such as person names. As a consequence, it is appropriate to find pairs of references similar with respect to their authors and the initial part of their title, as the rest of the reference text is more likely to contain errors or variations.

Table 21: Merging threshold and evaluation of its accuracy. Some datasets have borderline results, such as Icarus and the history of science, nevertheless results rapidly improve above the threshold, as false negatives disappear below it.

Statistic/Specialism	**A1-ec_hist**	**A2-hist_sci**	**B1-NIPS**	**B2-ACL**	**C1-icarus**	**C2-JCAP**	**D1-classics**	**D2-eng_lit**	**E1-neuron**	**E2-MBE**
Threshold	0.85	0.85	0.82	0.82	0.8	0.86	0.79	0.82	0.85	0.81
Precision 100 above	0.71	0.51	0.69	0.79	0.54	0.99	0.41	0.44	0.03	0.26
Precision 100 below	0.18	0.29	0.03	0.13	0.46	0.02	0.43	0.31	0.11	0.21

[22] Accuracy is the proportion of correct matches over the total considered. An accuracy of more than 0.5 above the threshold guarantees better than chance performance.

7.3 Method

Connectivity and giant component Expanding on the method introduced in Chapter 6, we measure: i) the proportion of connected components over the total possible (as before), and ii) the proportion of nodes in the giant component (Eq. 13), at steps in which we remove all edges below a certain weight threshold. Given an edge weight threshold t, we are thus interested in two measures, calculated at increasing t over networks B:

$$c(t) = \frac{C^t}{N} \tag{12}$$

$$g(t) = \frac{G^t}{N} \tag{13}$$

Where N denotes the number of publications, equivalent to the number of nodes in the network and also equal to the number of connected components in the disconnected network; C^t denotes the number of connected components after removal of edges with weight below t, as before; G^t denotes the number of nodes in the giant component after removal of edges with weight below t.

Core literature A complementary view on the granularity of topics in different specialisms can be given by considering the connectivity properties of the reference overlap bibliographic coupling network when removing highly cited sources (cf. hypothesis 6). The network will fragment after the removal of a proportion of highly cited sources, but it will do so at different speeds and times. Crucially, the more the specialism globally relies on shared sources (i.e. cited across topics), the less rapidly the network will initially fragment during such process; the more the specialism topically relies on core sources (i.e. cited within topics), the less rapidly the network will fragment once topics have been reached during such process. We compare two processes considering the directed citation network of a specialism: one where we remove increasing fractions of cited sources in reverse order by the number of citations they received (from high to low), another where we remove cited sources at random. We then construct the reference overlap bibliographic coupling network and inspect its connectivity properties at regular intervals, as done in the previous subsection.

7.4 Results

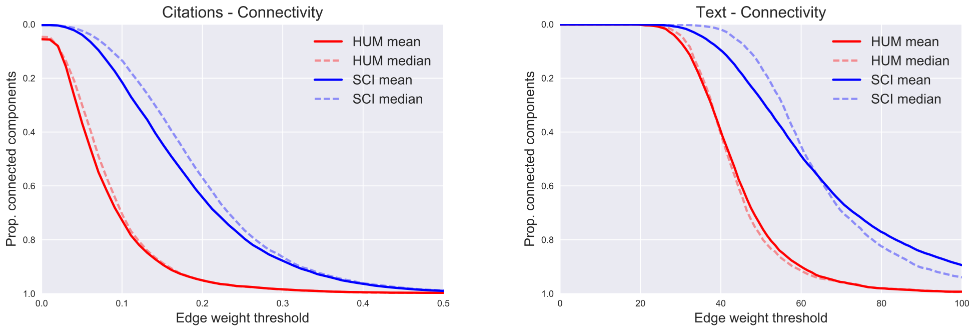
We start by providing an overview of the (reference overlap) bibliographic coupling networks of the ten specialisms under consideration, in Table 22, considering data covering 5 years: 2011 to 2015 included, hence the partial drop in the number of articles (nodes). It is worth noting that, under some aspects, specialisms in history and literature stand out. Namely, having fewer edges, a higher number of isolated nodes (with no edges), and an often lower density. This highlights from the very beginning that their networks are less well connected. Literature specialisms seem particularly weak in this respect, also considering their higher diameter, meaning that the distances in the network can be longer there.

Table 22: General statistics for the reference overlap networks of the ten specialisms under consideration.[a]

Statistic/Specialism	**A1-ec_hist**	**A2-hist_sci**	**B1-NIPS**	**B2-ACL**	**C1-icarus**	**C2-JCAP**	**D1-classics**	**D2-eng_lit**	**E1_neuron**	**E2_MBE**
# nodes	812	1652	1843	1410	2088	2834	1313	918	1508	1324
of which isolated	5	40	7	2	4	4	151	58	6	2
# edges	16'555	52'518	82'792	110'230	134'428	396'347	23'320	5456	51'589	114'803
density	0.05	0.038	0.049	0.111	0.062	0.099	0.027	0.013	0.045	0.13
diameter	5	7	5	5	5	6	15	11	6	5
global clustering	0.37	0.38	0.36	0.44	0.39	0.51	0.32	0.26	0.29	0.44

[a] Informal definitions: an isolated node is one without edges; the density of the graph is the proportion of existing edges over the maximum possible given the amount of nodes; the diameter is the longest shortest path between any two nodes in the graph; the global clustering of the graph is the number of existing triangles of nodes (or closed triplets) over the maximum possible. For formal definitions refer to Newman [2010]. Statistics were calculated using igraph [0.7.1] [Csardi and Nepusz, 2006].

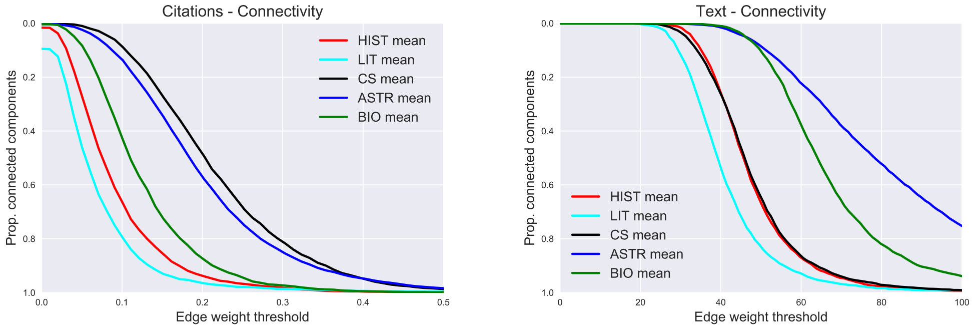
We compare next the reference and text similarity networks (built using Eq. 9 and 10 respectively, from Section 6.1). We group humanities specialisms (A and D) and science specialisms (B, C and E), following the hypothesis that the humanities specialisms are more rural, and science specialisms more urban. Consider first Equation 12. In Figure 33 we plot $c(t)$ on the y axis versus t on the x axis, averaging results over the humanities and the sciences, for both networks. In Figure 34 we give the same results, averaged over every specialisms instead. Consider next Equation 13. In Figure 35, we plot $g(t)$ on the y axis versus t on the x axis, with the same set-up as in Figure 33. Results for the size of the giant component are coherent with those for connectivity.



(a) Reference overlap.

(b) Textual similarity.

Figure 33: Mean and median connectivity of the reference (left) and text similarity (right) networks, grouped into the humanities (red/grey) and the sciences (blue/dark grey). Please note the y axis is reversed and goes from 1 to 0 instead.

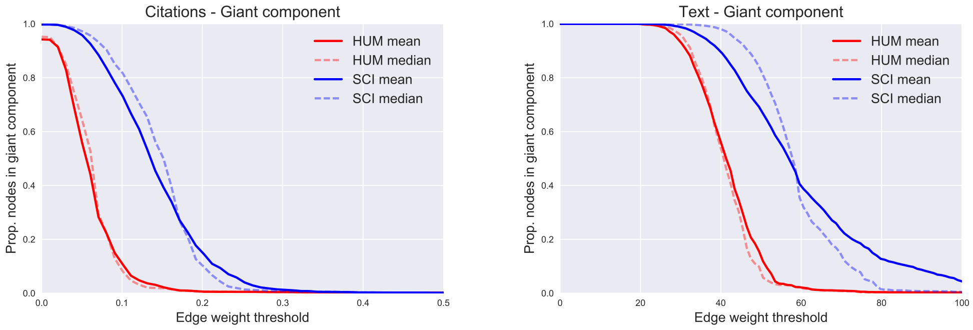


(a) Reference overlap.

(b) Textual similarity.

Figure 34: Mean and median connectivity of the reference (left) and text similarity (right) networks, by specialism. Legend: HIST: A, LIT: D, CS: B, ASTR: D, BIO: E. Please note the y axis is reversed and goes from 1 to 0 instead.

Our results clearly highlight a lower overall connectivity for specialisms in the humanities, both over reference and textual similarities. Individually, specialisms behave differently. Astrophysics (C) has both high reference and textual similarities, while computer science (B) has higher reference and lower textual similarity (identical to history), biology (E) has higher textual and lower reference similarity, being closer to history than astrophysics in this respect. Nevertheless, all scientific specialisms have higher connectivity than specialisms in the humanities, across both similarity measures, with history presenting slightly higher similarity than literature. This result indicates that research topics are finer-grained in the humanities than in the sciences, as discussed in hypothesis 1, and this is consistent both with respect to reference overlap and textual similarity over titles and abstracts. However, the substantial observed variety across specialisms indicates that their cognitive and social structures are likely not fully explained within the rural and urban analogy.



(a) Reference overlap.

(b) Textual similarity.

Figure 35: Mean and median size of the giant component of the reference (left) and text similarity (right) networks, grouped into the humanities (red/gray) and the sciences (blue/dark gray).

Given hypothesis 1 (i.e. urban specialisms maintain a higher connectivity, due to the presence of a lower number of larger topics), hypothesis 2 follows if urban specialisms also have more authors overall and more co-authorships, at the same weight threshold. Urban specialisms in our dataset indeed possess a higher number of unique authors (and a higher average number of co-authors per publication, cf. Table 20), thus hypothesis 2 follows immediately. This is because larger topics contain necessarily more authors than in rural specialisms, giving confirmation to the higher people-to-problem ratio of urban specialisms. In Table 23 we report the average and median size of the connected components with more than one node, and the corresponding people-to-problem ratio (number of unique authors per connected component), at different weight thresholds over the reference overlap network. Both the size of connected components and the people to problem ratio possess very skewed distributions. At the same time the size of topics and, especially, the people to problem ratio are sensibly higher for scientific specialisms, as expected.

Table 23: Size of connected components mean (median) and people-to-problem ratio mean (median) for the reference overlap networks, calculated at different thresholds considering components with more than one node. By people with consider unique authors active in the component. An author can be active in more components via multiple publications.

Specialism / Statistic	**Threshold t=0.1**		**Threshold t=0.2**		**Threshold t=0.3**	
	Topic size	**p-t-p**	**Topic size**	**p-t-p**	**Topic size**	**p-t-p**
A1 economic history	4.0(2)	5.9(4)	2.3(2)	3.5(3)	2.0(2)	2.9(3)
A2 history of science	7.7(2)	7.9(2)	2.8(2)	2.9(2)	2.4(2)	2.2(2)
B1 NIPS	93.4(2)	159.5(5)	6.6(2)	14.9(7)	3.0(2)	6.9(6)
B2 ACL	100.8(2)	164.4(6)	10.3(2)	20.3(6)	3.8(2)	9.2(6)
C1 Icarus	69.1(2)	164.8(10)	5.4(3)	19.0(11)	2.8(2)	9.6(7)
C2 JCAP	34.2(2)	79.1(6)	6.2(2)	15.6(6)	3.1(2)	7.0(5)
D1 classics	4.8(2)	4.4(2)	2.6(2)	2.3(2)	2.6(2)	2.4(2)
D2 English literature	3.4(2)	3.6(2)	2.3(2)	3.0(2)	2.2(2)	3.2(2)
E1 Neuron	14.8(2)	89.3(20)	3.0(2)	19(13)	2.5(2)	15.6(11)
E2 MBE	8.6(2)	35.2(12)	2.8(2)	9.6(7)	2.4(2)	9.1(5)

We further conducted the same experiments on all journals part of specialisms A and D individually (i.e. over networks of articles from the same journal only), to verify whether defining a specialism as an aggregation of articles from many journals would not artificially reduce the overall connectivity of the network. Indeed, no journal taken individually presents results markedly different from the overall trend of the respective specialism, thus we conclude that the lower overall connectivity in the humanities specialisms is not an artefact of journal aggregation. We omit these results for brevity.

The underlying process described by equations 12 and 13 is illustrated in Figure 36, for the case of specialisms B1 (NIPS) and D1 (Classics), considering increasing t (0.1, 0.2 and 0.3, left to right). It is possible to appreciate how the NIPS specialisms is not only denser at low t, but is also maintaining larger connected components at higher thresholds, according to hypothesis 1 and as detailed in Table 23.

Moving to consider the reliance of specialisms on their cited sources, we

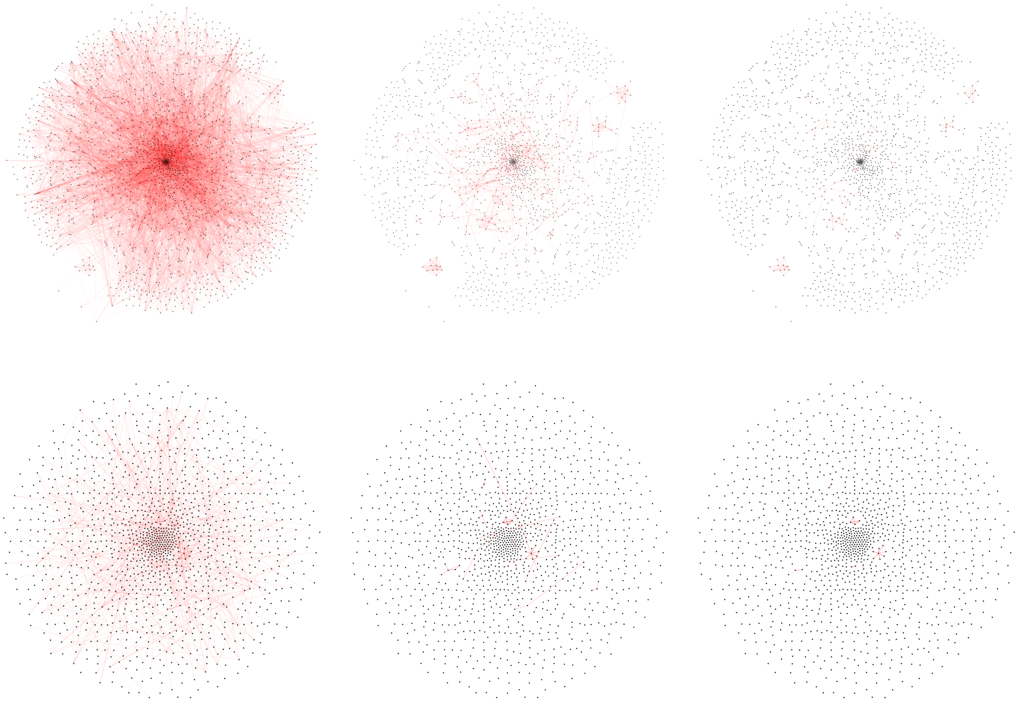
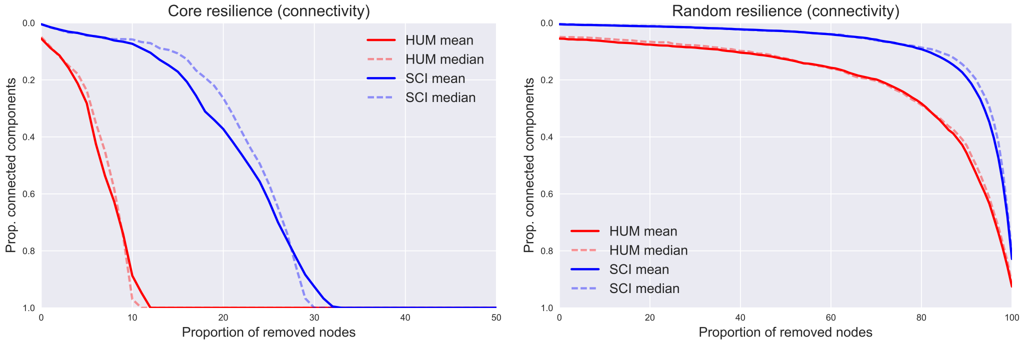


Figure 36: Illustration of two reference overlap networks during the process of removal of edges below a given threshold. The same network layout is kept in all figures (using Force Atlas 2 in Linlog mode from Gephi 0.9.1 [Bastian et al., 2009; Jacomy et al., 2014]). Above: B1 NIPS. Below: D1 Classics. Thresholds: 0.1 (left), 0.2 (centre), 0.3 (right). The NIPS network presents several connected components even at relatively high thresholds, while the Classics network becomes almost disconnected already at threshold 0.2.

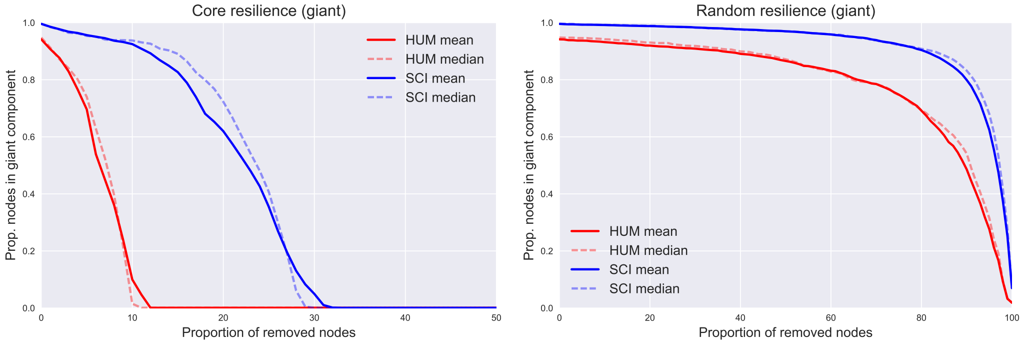
show results in Figure 37 for connectivity and Figure 38 for the giant component, averaging as before over the humanities and the sciences. In both cases, a process of removal in order of received citations is compared with one where edges where removed at random.



(a) Reference overlap.

(b) Textual similarity.

Figure 37: Connectivity of the reference similarity networks to the removal of highly cited sources first (left) or at random (right), divided in the humanities (red/grey) and the sciences (blue/dark grey). The proportion of removed nodes is in %, thus 10 means 10%. Please note the y axis is reversed and goes from 1 to 0 instead.



(a) Reference overlap.

(b) Textual similarity.

Figure 38: Changing size of the giant component of the reference similarity networks to the removal of highly cited sources, divided in humanities (red/grey) and sciences (blue/dark grey). The proportion of removed nodes is in %, thus 10 means 10%.

Two results emerge. Firstly, scientific specialisms are more reliant on core literature both at the specialism level and at the topic level, witness the higher resilience of their bibliographic coupling networks to the removal of the core literature at all stages. Secondly, the humanities present less well-connected bibliographic coupling networks in general, as shown by their lower connectivity apparent at all stages of the random removal process. To be sure, this phenomenon is much stronger in literature than history. We might conclude that the humanities possess a more fragmented intellectual base in terms of reference overlap, and in particular share fewer core sources at the specialism level, which in part contradicts what stated in hypothesis 6. The effect of the removal on the size of the giant component also suggest that the core literature is more substantially shared at the specialism level in the sciences, and not just within larger topics. These results are even stronger for literature, a purely humanities field, than history, which at times borders the social sciences – and especially so economic history. If rural specialisms are fragmented into small topics, as shown, and at the same time possess fewer core sources, it follows that their fragmentation is in part due to more focused, topic-specific or unique, non-overlapping referencing behavior. But also within topics the connectivity is higher in the science specialisms.

7.5 Discussion

The aim of this chapter was to cast a comparison across specialisms traditionally considered part of the sciences or the humanities, and among them history. We did so by offering a possible quantitative operationalization of Becher and Trowler [2001]'s conceptualization of the social and cognitive structure of research specialisms as rural or urban. According to this conceptualization rural specialisms show, in comparison to urban ones, (1) a higher number of smaller topics being researched, (2) a lower people-to-problem ratio, (3) longer publications, that (4) contain more references, and also show (6) to share comparatively more references across topics than within, and (8) to have less co-authorships.

We have proposed an operationalization of this conceptualization of the social and cognitive structure of research specialisms by comparing the textual and reference connectivity among publications within ten humanities and science specialisms. We used publication venues (journals) to proxy specialisms and well-connected clusters in the bibliographic coupling network of publications to proxy topics. Considering reference connectivity first, and focusing on hypothesis 1 and 6, we found that hypothesis 1 is confirmed at the specialism level, as science specialisms are overall better connected than in the humanities, with some disciplinary variations. Similarly using textual connectivity, we saw a stronger connectivity in the sciences, especially at smaller topics, likely the effect of technical jargon and a higher degree of specialization. An exception at the textual level is computer science, behaving on par with history. With respect to hypothesis 2 we found strong supporting evidence for a considerably higher people-to-problem ratio in urban specialisms, or the number of active authors per topic. However, topics are not that easily defined and we did not find many distinct clusters in any specialism. This leads us to argue that within the sciences, specialisms are comparatively well-connected at both the level of general, larger topics and the level of smaller, tighter ones, but that the distinction between these two levels of the cognitive structure is not as clear as Becher and Towler suggest. Within the humanities we find a comparatively lower connectivity also on the level of the specialism. This means that we do not find any evidence for the idea that humanities scholars tend to cite more broadly to establish an intellectual base for their contribution within the specialism as a whole.

In light of these findings we suggest to re-evaluate the use of the rural versus urban conceptualization. Despite the fact that some already established elements of this conceptualization find confirmation in our analysis (e.g. length of publications, number of active authors, people-to-problem ratio), the specific cognitive structure of specialisms and topics is only partially reproduced. Rather, we find that science specialisms show an overall cohesion that suggests that scholars work in a particular paradigm in which topics are not necessarily clearly distinguished. The overall fragmentation of the humanities specialism suggests instead a less unified cognitive structure, at least to the extent to which this is articulated through reference lists and textual similarity. It is still possible that in these humanities specialisms a particular paradigm is dominant without scholars having to articulate it or having to make reference the historical sources that lay at the basis of this paradigm. Despite this, it should not be the case that the relatively limited size of the corpus (both in number of articles and over time) limits our ability to find evidence for particular topics that might exists across a broader spectrum of publications, since the choice of journals was made in order to be representative of the work being published in a specialism. Finally, and importantly, the journal article might assume different roles in the humanities, for example as a more specialized form of publication, as monographs play a more important role there than in the sciences [Thompson, 2002; Williams et al., 2009]. Going back to the open question related to Fuchs' theory of scientific organization from Chapter 6, we still need to discuss if history is intellectually specialized or fragmented as a discipline. The results of this chapter would suggest to favor fragmentation, as the very limited presence of core works is certainly evident in all specialisms in the humanities, history included. This was indeed the main result which was not in agreement with Becker and Towler's conceptualization.

8 Conclusions

At the beginning of the thesis we advanced four overarching questions to guide our work:

1. Can we enlarge the bibliometric data coverage for the humanities?
2. How can we bibliometrically represent and study research fields in the humanities, with a focus on their intellectual organization and informed by a theoretical framework?
3. Can we individuate structural elements in the intellectual organization of fields of research?
4. How is new knowledge accumulated within such intellectual organization(s) and how is this changing over time?

In this last chapter we review these questions and summarize our main findings. We then attempt a characterization of the academic discipline of history as a way to create knowledge on the human past, and we conclude by discussing some limitations and directions for future work.

The first question raises an important issue in bibliometrics: the poor data coverage for the humanities. In Chapter 3 we proposed to rely on research libraries in order to acquire representative corpora of publications which can then be mined for citations. A specific case study on the history of Venice was taken as an opportunity to develop a more general approach to build a citation index for the humanities, which includes all publication typologies and all cited sources. The proposal is to follow the same approach that allowed libraries to build large catalogs: the collaborative division of work. We thus proposed and developed a platform made of two applications: a digital library, which can be used by any library and embeds a citation extraction pipeline, and a citation index which federates all citation data into a unique database. This solution has another important benefit: the indexation of citations to primary sources would interconnect library collections with other repositories such as archives, museums and other libraries, providing for an information retrieval engine of a potentially broader interest for the scholarly community and beyond. The answer to the first question is thus positive, yet there is a long way to go in terms of further developing the technical infrastructure and fostering the collaborative effort before a citation index for the humanities can reach momentum and scale. Eventually, the case study allowed to produce new citation data which was in part used in the present work.

The following Chapter 4 focused on a novel dataset produced as part of the Venetian case study: a book to book citation dataset where the citing books were relatively recent, mostly published from the 1980s onwards. The research front, or anyway the intellectual organization of the citing publications was thus represented via their bibliographic coupling network, which highlighted how the specialism is broadly organized by disciplinary areas (history, art history) and chronological periods. Furthermore, the core, highly cited sources belonged to two groups: 1) primary sources (e.g. edition of documents) and reference works (e.g. catalogs or repertories), and 2) scholarly monographs. Works in the former group, in particular, were found to be potentially very old (19^{th} century or before), highlighting how the influence of a work of this kind can unfold over very long periods of time. The two groups can be roughly qualified as works on evidence, where primary sources or relevant information are made available to further the research efforts of other scholars, and works on interpretation. To be sure, this distinction is blurred by the fact that any work in historiography relies on evidence and requires some interpretive effort, a typical example being critical editions of documents, yet it is worth distinguishing the two categories by the purpose they have as scholarly outputs. Works on evidence have the main purpose of facilitating the work of other members of the community, works on interpretation have the main purpose of directly contributing to the community's body of knowledge. The crucial importance of works on evidence for the historians of Venice, also confirmed for other specialisms in Chapter 5, led us to suggest they act as "research technologies" in other disciplines do with respect to their influence on the intellectual organization of the specialism.

The core sources emerged in Chapter 4 as a key structural element in the intellectual organization of the specialism history of Venice, mapped as a bibliographic coupling network. In Chapter 5 we thus sought to measure their influence on such a representation, and study it with three case studies/datasets of varying size and scope. We saw that overall the core, always mostly composed of books where applicable, can play two roles in connecting the bibliographic coupling network of a specialism or a field: connect locally within communities and connect globally across them. A global action, which we compared to the weak ties of social networks, is performed mostly by books which sometimes emerge beyond the scope of a small community and acquire popularity across several ones. In this chapter we also confirmed the fundamentally multilayer character of history's intellectual organization when considering both citing publications and cited sources. In particular, we confirmed that books and journal articles can play different, possibly complementary roles as cited sources, thus with respect to the scope of the contributions they contain.

The second part of the thesis was instead more directly informed by two theoretical frameworks from the sociology of science: Fuchs' theory of scientific change and Becker and Trowler's characterization of academic fields as rural and urban. Chapter 6 posed the question if history, represented by a set of specialisms therein, is becoming more or less fragmented over time. This question was explored by considering the connectivity property of two bibliographic coupling networks: one constructed using reference overlap, another using text similarity. Results clearly pointed to a declining connectivity of the former network, and a stable outlook for the latter, suggesting how most specialisms in history might be gradually drifting apart in terms of shared references, but maintaining cohesion in terms of language used and perhaps even topics and ideas discussed. A notable exception was economic history, where co-authorships are significantly on the rise and where the connectivity of both networks is instead stable. In the last Chapter 7 we instead operationalized Becker and Trowler's characterization and tested it on a collection of specialisms from history, literature, biology, physics and computer science. Our results pointed to a marked difference between so called rural specialisms (the humanities) and urban ones (the sciences) when considering both reference overlap and text similarity networks. Furthermore, despite the fact that Becker and Trowler's framework was found to be mostly accurate in distinguishing among the two groups at a coarse level, it was not predictive on a key element: the reliance on core sources. In fact, rural specialisms were found to possess proportionally fewer core sources, and to be more fragile at the level of the whole specialism: a tiny fraction of works was all that connected the whole specialism across smaller topics of research. Conversely, urban specialisms were found to possess more core sources shared across topics, and to be overall better connected over both references and language. Recovering Fuchs' theory, we were thus able to suggest that history is an intellectually fragmented more than specialized discipline, as the mutual dependence among scholars is very low and akin to rural areas of research.

8.1 A characterization of history

We can now advance a more coherent characterization of history, mostly in the form of hypotheses, as a discipline producing a certain form of narrative knowledge of the past. First of all, perhaps not surprisingly, there is a multilayered system of publication typologies in history, as well as a parallel system of cited sources. It is likely that different publication typologies serve distinct and complementary purposes, as it was shown for books and journals articles in the study of core sources. It is therefore equally likely that a variety of scholarly profiles exist among historians, according to different preferences in terms of outputs, audiences, methods and interests. This variety contributes to a multifaceted intellectual organization in ways that are still largely to be explored.

History is furthermore a rural discipline, intellectually fragmented and advancing at a slow pace, at least judging by the age of cited literature, where scholars possess a low mutual dependence and are little keen on collaborating. This profile aligns with Price's characterization of disciplines which grow 'from the body', requiring long time to digest previous literature which ages slowly in turn, as it was seen multiple times in the present study [De Solla Price, 1970; Cozzens, 1985]. A very similar characterization was found to apply to literary studies too [Hammarfelt, 2012a]. All these are traits typically attributed to the humanities. It is worth reiterating again that one of the few exceptions we found, economic history, is not by chance closer to the social sciences and showing, among other characteristics, a rising propensity for collaborations over time. A trait which is perhaps peculiar to history, at least to the best of our knowledge, is the importance of reference works at least within specialisms (such as it was shown for the history of Venice and the history of the book). If we consider how knowledge accumulates and thus how the intellectual organization of a specialism in history looks like, reference works and work on primary sources might play a different role than more mainstream interpretive work.

We previously advanced a distinction between works focusing on evidence and on interpretation. A focus on *evidence* characterizes scholarly work which improves the retrieval, access and further use of evidence or information on it by other scholars. Examples are catalogs or edition of sources, which are meant to aid scholars find materials and assess them, or to work on some evidence at the aid of a preliminary effort made by someone else. This kind of work thus requires shared criteria for assessing its quality. A focus on *interpretation* requires instead less commitment in terms of shared criteria, as it is mostly evaluated in terms of its novelty and originality. A work of interpretation does not as much provide something to directly rely upon as instead something to enrich our current understanding of a phenomenon. To be sure, this distinction in practice is not as sharp as we made it sound here, as every work on evidence requires interpretation and any work of interpretation requires a sound grounding in primary evidence. Recovering Whitley's framework, discussed in Section 2.6.1, we may suggest that the intellectual organization of historians might change according to their focus. Work on evidence might be conducted in the presence of a lower technical task uncertainty and higher functional dependence among scholars, who share more in terms of questions, methods and criteria of quality, thus enabling for a somewhat coordinated production of knowledge, despite its persisting specific scope.

Another distinction we advance here has to do with the exploratory strategies scholars adopt when faced with the question of how to select and approach a research project. An *in-depth exploration* favors taking something and bettering or directly improving the existing art on it. This approach is very common in fields with high mutual dependence among scholars and low task uncertainty, using Fuchs' terminology, such as computer science. At the same time, this approach is quite uncommon in history or the humanities more generally, yet examples exist (e.g. multiple critical editions of the same source, or gradual improvements to catalogs and inventories). A second way to explore is *in-breadth*, by finding novel topics, sources, methods, ideas to explore, or mixing old ones in novel ways, instead of directly reconsidering existing ones. An in-breadth approach is more common in history, especially given the scope of topics and sources which can be used.

We argue these modes of focus and exploration have specific consequences on the intellectual organizations of specialisms in history, given the disciplinary traits we listed above, especially via the accumulation of novel literature. We sketch these consequences in Figure 39. The most common scenario is interpretive work using an in-breadth approach, which leads to intellectual fragmentation given the substantial lack of strong unifying theories or methods in history. Alternatively, in-depth works of interpretation can attempt to reconsider existing results and directly improve on them, or anyway proceed more gradually, with the result of a rise in specialization as the intellectual organization should become locally stronger in this case (high mutual dependence of a small part of the landscape, where the specific work is relevant).

If we move over to a focus on evidence, an in-breadth approach entails exploring new projects leading to local specialization too, as scholars will use these results in small areas of the landscape, while an in-depth one entails bettering existing results instead. This latter kind of contribution should rise the concentration of citations to the given work, gradually creating a sort of hub in the landscape.

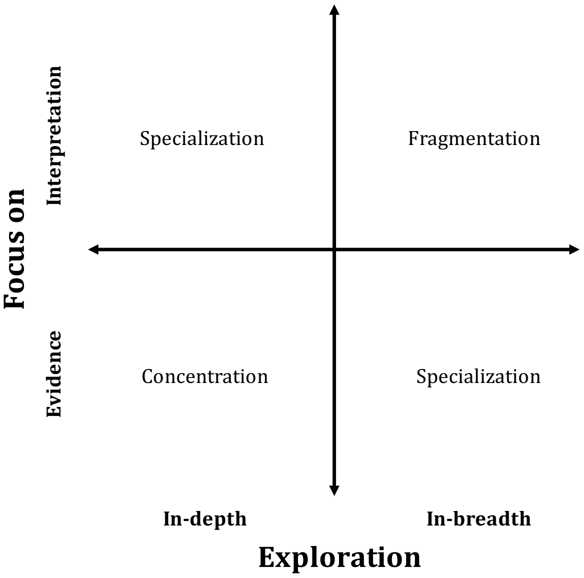


Figure 39: The effects on the intellectual organization of a specialism in history according to in-depth or in-breadth exploratory strategies and a focus on interpretation or evidence.

In Figure 40 we show some examples of scholarly outputs and how they might map on the hypothetical plane we just proposed. Evidently, these positions are hypothetical and do not entail that individual works cannot fare differently (e.g. a monograph mostly exploring in-depth). Scholarly monographs likely represent the most interpretive and in-breadth kind of output, while journal articles possibly represent a more specialized form of publication, containing focused results more explicitly grounded in evidence. This is what sometimes historians call technical works. On the side of evidence, we might (somewhat arbitrarily) distinguish among outputs with a more or less marked interpretive component (e.g. critical editions and catalogs), but especially among outputs which explicitly focus on bettering or expanding existing things (in-depth), or creating new ones (in-breadth). Bettering existing things should further strengthen their importance for the community and thus help concentrate citations, conversely novel things might require further specialization over a smaller area of the intellectual landscape. To be sure, the degree of integration of a reference work is quite important in this setting: for example, if a new database abides to standards and is easily integrated with existing technologies, its effect might more rapidly move from specialization to concentration.

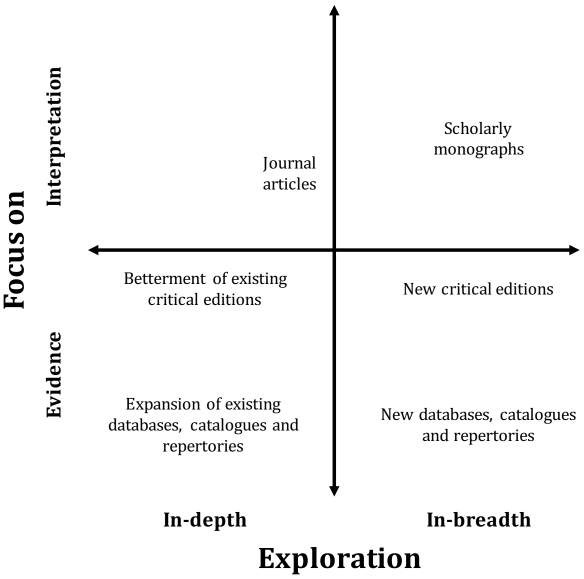

Focus on
Interpretation
Evidence
Journal articles
Scholarly monographs
Betterment of existing critical editions
New critical editions
Expansion of existing databases, catalogues and repertories
New databases, catalogues and repertories
In-depth
In-breadth
Exploration

Figure 40: A hypothetical collocation of some scholarly outputs in history according to their proposed impact on the intellectual organization of a specialism.

We can now reconsider what might be the role of humanities computing, where we mean the part of the digital humanities more closely involved with work on evidence and the application of computational tools and methods to the traditional questions and objects on inquiry of the humanities. First of all, humanities computing fosters a transition in the humanities from the analogue to the digital in terms of information technologies. This is not trivial as the humanities are disciplines of the book: their information technologies and modes of ordering, sorting, indexing and searching information are largely those developed for books. Examples are indexes and footnotes. This gradual transition might influence the intellectual landscape of these disciplines by changing their modes and forms of publication and organization of the body of knowledge.

A second area where the role of humanities computing is increasingly more felt is the social organization of these disciplines, with important consequences on their intellectual one too. This approach to the digital humanities fosters a 'long tradition of patter searching that exists in the humanities and sciences alike' [Bod, 2018, 24]. Work on evidence or interpretative work using quantitative approaches, such that is possible once digital evidence has been accumulated, leans towards a social organization of research which is different from the traditional fragmented one of the humanities, and closer to the model of the *laboratory*. This organization unit is composed of several persons, working in teams on a variety of projects, each with specialized knowledge and often complementary technical skills. In a laboratory there exists a division of labor, contrary to when the scholar works alone. This mode of organization might have the following further consequences: the intermediation of technology and of other scholars' work (cumulative research) will rise in importance, especially with works on evidence; the level of codification of the knowledge produced will rise too, as will the reliance on formalization; the laboratory is a fund-hungry entity, thus larger projects with more generous funding will become more important; lastly, there will be a technicization of the humanist, in terms of skills and of domain knowledge. All these changes, if they will happen, should gradually drift the humanities away from a rural organization towards a more urban one.

These potential changes can again be put in the context of the theoretical frameworks previously discussed. Whitley's classification of the humanities as fragmented adhocracies, which largely applied to history too, would gradually shift towards a professional adhocracy, where the technical task uncertainty lowers as the functional dependence among scholars rises. To the extent that shared research technologies become increasingly important and not specific to a small community or specialism, a further transition might happen towards technologically integrated bureaucracies, where knowledge integration and low strategic task uncertainty is achieved through technologies instead of theory. Another perspective is given in Fuchs [1993b], where three epistemologies of scholars are described: pragmatism, positivism and hermeneutics. Positivism is essentially the epistemology guiding Kuhnian normal sciences, and hermeneutics instead the epistemology of conversational, textual and decentralized fields. Historiography which aims at reconstructing the past using essentially philological methods on archival materials, takes a positivist epistemology. Historiography trying to tie the present and the past and convey reflections through narratives adopts instead an hermeneutics one. The third option is the epistemology of research fronts: areas of research too young or innovative to adopt anything but a 'flexible entrepreneurial pragmatism' [Fuchs, 1993b, 30]. Interestingly, Fuchs considers Big Science projects as those where pragmatic considerations, such as politics and return on investment, are most felt. An increasing adoption of a pragmatic epistemology and the laboratory organization in the humanities might happen, or is indeed already taking place due to the digital turn. As a consequence, the digital humanities might bring profound changes in both the socio-institutional and intellectual organizations of the humanities. To what extent this process will be profound and desirable remains to be seen.

8.2 Limitations and future work

This work has some limitations worth highlighting. First of all, its scope is limited not only because we only considered history, but also because we made some choices within history on which specialisms and datasets to use. Nevertheless, despite the fact that further work might very well improve the picture, it seems possible to at least consider our conclusions as representative and plausible, given the way such selection choices were made and justified. To be sure, the availability of data quite sensibly constrained our options in this respect, so much so that this must be considered as a second limitation of the thesis, albeit one were we contributed towards to some extent. A third significant limitation is due to the challenging operationalization of the theories we used, borrowed from the sociology of science. It must be stressed that all these theories lack mathematical formalization and often use ill-defined concepts. For example, two key concepts from Becker and Trowler, specialisms and topics, are assumed to be defined identically for both rural and urban fields in order to allow for comparisons, yet no indication on how to do so is provided.

Furthermore, a last limitation rests on our methodological choices: the widespread use of publication data such as citations and texts, and of network representations could be very well complemented by alternative approaches. In particular, the use of texts was quite limited and could have been expanded to include comparisons at the level of topics (in the topic modeling sense) and not just vocabularies. A question should in particular be posed here:

was the choice of methods appropriate to the task at hand, and is it indeed possible to study the intellectual organization of a field of research mainly through citations? We reply positively. This work has shown that, provided the necessary citation coverage in scale, time and depth, it is indeed possible to characterize the intellectual organization of history, and by extension the humanities. These requirements exist in order to preserve the granularity of a phenomenon which unfolds slowly and through complex interactions, such is knowledge accumulation in the humanities. To what extent hidden, sometimes impossible to measure influences such as tacit knowledge can influence or even explain such a structure and its dynamics remains difficult to say [Crane, 1975; Polanyi, 1997]. This brings us back to the original distinctions between the sciences and the humanities which were discussed in the introduction: citation maps can convey the current configuration of a system, they can be used to monitor and understand what is happening, much less why, and cannot in themselves explain without reliance on theory. Despite these limitations, and possibly other ones we did not mention, we believe this thesis to be a step in a set of useful directions: enlarging the bibliometrics data coverage of the humanities, connecting theoretical results in the sociology of science with empirical bibliometrics research, and eventually expanding our understanding of the humanities as academic disciplines at a crucial time of change.

Many directions for future work lay open, we list here a few:

- The indexation of the literature and the sources published or used in the humanities. Citation indexation is perhaps an obvious goal. Despite the fact that what we know of the humanities so far clearly points to the fact that citation data might be less useful in the humanities for research evaluation, this endeavor retains all its usefulness from the information retrieval and mapping points of view. Yet the indexation of publications and sources can be done in other complementary ways than using citations, and this is a key area of future development for the digital humanities too.

- As we broaden the quality and quantity of data, we should in parallel expand our methods. With respect to citations, the humanities should be considered as a multilayer system and the context of citations should be better explored and understood. It is well know that humanists possess a rich "citation vocabulary" which should not be ignored. Furthermore, the humanities might perhaps be better understood using the full text of publications as a way to complement anyway fragmented citation networks: this is in general a still little explored area of research in bibliometrics.

- Thirdly, a bibliometrics point of view might contribute to assess the impact of the digital humanities within the humanities. As we found in this work, historians have been relying on reference works and editions of sources since a long time, and this is indeed a key area of contribution for the digital humanities and humanities computing before them. The impact of these contributions could be studied bibliometrically too. Furthermore, the digital activities of humanists might be object of investigation as well: both their critical reflection on digital media and their broadening use of novel forms of communication.

- Lastly, as we start to accumulate a certain amount of studies considering individual disciplines or specialisms, the time might soon be ripe for a systematic and large scale comparative analysis of all disciplines part of the humanities from a bibliometrics perspective informed by the sociology of science. The digital turn brings an urgency to map the past and current intellectual and social organization of all sciences, the humanities in particular given how little we know of them, in order to understand and control future developments.

This thesis humbly rests on the shoulders of several areas of research which interconnect not nearly enough: it is expression of an interest in history from a bibliometrics perspective, informed by the sociology of science, with an eye on digital infrastructure for the humanities and the recent trends unfolding in the digital humanities. Perhaps the key goal for future work should be to keep fostering interconnections, as knowledge knows no boundaries but those posited by our imagination.

9 Reproducibility: code and data availability

The work on the citation indexing of the scholarly literature on the history of Venice is still ongoing. Nevertheless, a number of code and data releases have already been made:

- A corpus of citations between books, based on the references extracted from the reference lists of 700 monographs, as detailed in Section 3.2. See Romanello and Colavizza [2017] and `http://doi.org/10.5281/zenodo.377047`.
- A corpus of more than 40'000 annotated references and the code to train reference mining CRF models using them [Colavizza and Romanello, 2017], available at `https://github.com/dhlab-epfl/LinkedBooksReferenceParsing` and `http://doi.org/10.5281/zenodo.579679`.
- A deep learning architecture for reference mining, using the same dataset [Rodrigues Alves et al., 2018], available at `https://github.com/dhlab-epfl/LinkedBooksDeepReferenceParsing`.

The two interfaces of the Scholar Index can be accessed as follows. Digital library: `https://library.venicescholar.eu` (requires login); citation index (Venice Scholar): `https://venicescholar.eu`. Try search for the historian "Patricia Fortini Brown", for example. The project Scholar Index also have a website, where it is possible to track its progress: `https://scholarindex.eu`.

Finally, all experiments and plots for Chapter 4 are available at `https://github.com/Giovanni1085/core_literature_historians_venice`. The remaining Chapters 5, 6 and 7 use proprietary data which could not be released.

Funding

This research has been funded by the Swiss National Fund with grants 205121_159961 and P1ELP2_168489. It was also indirectly supported by the École Polytechnique Fédérale de Lausanne and the Lombard Odier Foundation thanks to their funding of the Venice Time Machine project.

References

P. Ahlgren, P. Pagin, O. Persson, and M. Svedberg. Bibliometric analysis of two subdomains in philosophy: free will and sorites. *Scientometrics*, 103 (1):47–73, 2015.

J. Ardanuy. Sixty years of citation analysis studies in the humanities (1951-2010). *Journal of the American Society for Information Science and Technology*, 64(8):1751–1755, 2013.

S. Balietti, M. Mäs, and D. Helbing. On disciplinary fragmentation and scientific progress. *PloS one*, 10(3):e0118747, 2015.

A.-L. Barabási. *Network science*. Cambridge University Press, Cambridge, 2016.

A. Barrett. The information-seeking habits of graduate student researchers in the humanities. *The Journal of Academic Librarianship*, 31(4):324–331, 2005.

M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media*, 2009.

T. Becher. *Academic Tribes and Territories: intellectual enquiry and the cultures of disciplines*. Open University Press, Buckingham, 1989.

T. Becher and P. Trowler. *Academic tribes and territories: intellectual enquiry and the culture of disciplines*. Open University Press, Philadelphia, PA, 2nd edition, 2001.

V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

R. Bod. *A new history of the humanities: the search for principles and patterns from Antiquity to the present*. Oxford University Press, Oxford, 2013.

R. Bod. Has There Ever Been a Divide? A Longue Durée Perspective. *History of Humanities*, 3(1):15–25, 2018.

A. Bonaccorsi, C. Daraio, S. Fantoni, V. Folli, M. Leonetti, and G. Ruocco. Do social sciences and humanities behave like life and hard sciences? *Scientometrics*, 112(1):607–653, 2017.

L. Bornmann and H.-D. Daniel. What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1):45–80, 2008.

L. Bornmann and R. Mutz. 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, 2015.

M. Bourdeau. Auguste Comte. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2015 edition, 2015.

J. Bouterse and B. Karstens. A Diversity of Divisions: Tracing the History of the Demarcation between the Sciences and the Humanities. *Isis*, 106 (2):341–352, 2015.

K. W. Boyack, R. Klavans, and K. Börner. Mapping the backbone of science. *Scientometrics*, 64(3):351–374, 2005.

K. W. Boyack, D. Newman, R. J. Duhon, R. Klavans, M. Patek, J. R. Biberstine, B. Schijvenaars, A. Skupin, N. Ma, and K. Börner. Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches. *PLoS ONE*, 6(3):e18029, 2011.

G. Buchanan, S. J. Cunningham, A. Blandford, J. Rimmer, and C. Warwick. Information seeking by humanities scholars. In *International Conference on Theory and Practice of Digital Libraries*, pages 218–229. Springer, 2005.

V. Bush. As We May Think. *The Atlantic*, Feb. 1945.

K. Börner. *Atlas of science: visualizing what we know.* MIT Press, Cambridge, Mass., 2010.

K. Börner and A. Scharnhorst. Visual conceptualizations and models of science. *Journal of Informetrics*, 3(3):161–172, 2009.

K. Börner, C. Chen, and K. W. Boyack. Visualizing knowledge domains. *Annual review of information science and technology*, 37(1):179–255, 2003.

E. H. Carr. *What is history? The George Macaulay Trevelyan lectures delivered at the University of Cambridge January-March 1969.* Vintage, New York, 1998.

C. Chen. Science Mapping: A Systematic Review of the Literature. *Journal of Data and Information Science*, 2(2):1–40, 2017.

C. Chen and M. Song. *Representing Scientific Knowledge*. Springer International Publishing, Cham, 2017.

P.-S. Chi. Differing disciplinary citation concentration patterns of book and journal literature? *Journal of Informetrics*, 10(3):814–829, 2016.

G. Colavizza. The Core Literature of the Historians of Venice. *Frontiers in Digital Humanities, Digital History*, 4(14), 2017a.

G. Colavizza. The structural role of the core literature in history. *Scientometrics*, 113(3):1787–1809, 2017b.

G. Colavizza. Understanding the history of the humanities from a bibliometric perspective: Expansion, conjunctures and traditions in the last decades of Venetian historiography (1950-2013). *Forthcoming in History of Humanities*, 2018.

G. Colavizza and F. Kaplan. On Mining Citations to Primary and Secondary Sources in Historiography. In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015 3-4 December 2015, Trent*, pages 94–99, 2015.

G. Colavizza and M. Romanello. Annotated References in the Historiography on Venice: 19th–21st centuries. *Journal of Open Humanities Data*, 3, 2017.

G. Colavizza, M. Romanello, and F. Kaplan. The references of references: a method to enrich humanities library catalogs with citation data. *International Journal on Digital Libraries*, 18:1–11, 2017.

G. Colavizza, K. W. Boyack, N. J. van Eck, and L. Waltman. The Closer the Better: Similarity of Publication Pairs at Different Cocitation Levels. *Journal of the Association for Information Science and Technology*, 69(4): 600–609, 2018a.

G. Colavizza, T. Franssen, and T. van Leeuwen. An empirical investigation of the Tribes and their Territories: are research specialisms rural and urban? *Submitted to Journal of Informetrics*, 2018b.

G. Colavizza, M. Romanello, M. Babetto, V. Barbay, L. Bolli, S. Ferronato, and F. Kaplan. The Scholar Index: Towards a collaborative citation index for the Arts and Humanities. In *Proceedings of the Digital Humanities Conference*, 2018c.

S. Cole. The hierarchy of the sciences? *American Journal of Sociology*, 89 (1):111–139, 1983.

R. Collins. *Conflict sociology: toward an explanatory science.* Academic Press, New York, 1975.

R. Collins. Why the Social Sciences Won't Become High-Consensus, Rapid-Discovery Science. *Sociological Forum*, 9(2):155–177, 1994.

M. Coscia and M. Schich. Exploring Co-Occurrence on a Meso and Global Level Using Network Analysis and Rule Mining. In *MLG*, San Diego, 2011.

I. G. Councill, C. L. Giles, and M.-Y. Kan. ParsCit: an Open-source CRF Reference String Parsing Package. In *LREC*, 2008.

S. E. Cozzens. Using the archive: Derek Price's theory of differences among the sciences. *Scientometrics*, 7(3-6):431–441, 1985.

D. Crane. *Invisible colleges: diffusion of knowledge in scientific communities.* University of Chicago Press, Chicago London, 2. impression edition, 1975.

G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL `http://igraph.org`.

J. Cullars. Citation characteristics of monographs in the fine arts. *The Library Quarterly*, 62(3):325–342, 1992.

N. S. Davidson. "In dialogue with the past": Venetian Research from the 1960s to the 1990s. *Bulletin of the Society for Renaissance Studies*, 15(1): 13–24, 1997.

N. De Bellis. *Bibliometrics and citation analysis: from the Science citation index to cybermetrics.* Scarecrow Press, Lanham, Md, 2009.

D. De Solla Price. Networks of Scientific Papers. *Science*, 149(3683):510–515, 1965.

D. De Solla Price. Citation Measures of Hard Science, Soft Science, Technology, and Nanoscience. In C. E. Nelson and D. K. Pollock, editors, *Communication Among Scientists and Engineers*, pages 3–22. Heath Lexington Books, Lexington Mass., 1970.

D. De Solla Price. *Little Science, Big Science... and Beyond.* Columbia Univ. Press, New Yock, 1986.

A. A. Diaz-Faes and M. Bordons. Making visible the invisible through the analysis of acknowledgements in the humanities. *Aslib Journal of Information Management*, 69(5):576–590, 2017.

W. Dilthey. *Einleitung in die Geisteswissenschaften.* Duncker & Humblot, Leipzig, 1883.

J. Dupré. The Disunity of Science. *Mind*, XCII(367):321–346, 1983.

E. Dursteler. A brief survey of histories of Venice. In E. Dursteler, editor, *A Companion to Venetian History, 1400-1797*, pages 1–24. Brill, Leiden, 2013.

C. Emmeche, D. B. Pedersen, and F. Stjernfelt, editors. *Mapping frontier research in the humanities.* Bloomsbury Academic, London ; New York, 2016.

T. C. E. Engels, T. L. B. Ossenblok, and E. H. J. Spruyt. Changing publication patterns in the Social Sciences and Humanities, 2000–2009. *Scientometrics*, 93(2):373–390, 2012.

J. A. Evans. Electronic Publication and the Narrowing of Science and Scholarship. *Science*, 321(5887):395–399, 2008.

D. Fanelli and W. Glänzel. Bibliometric Evidence for a Hierarchy of the Sciences. *PLoS ONE*, 8(6):e66938, 2013.

P. Feyerabend. *Against method.* Verso, London; New York, 4th edition, 2010.

T. Finkenstaedt. Measuring research performance in the humanities. *Scientometrics*, 19(5-6):409–417, 1990.

S. Fortunato and M. Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.

S. Fortunato and D. Hric. Community detection in networks: A user guide. *Physics Reports*, 659:1–44, 2016.

T. Franssen and P. Wouters. Science and its significant other: Representing the humanities in bibliometric scholarship. *arXiv preprint arXiv:1710.04004*, 2017.

J. Fry and S. Talja. The intellectual and social organization of academic fields and the shaping of digital resources. *Journal of Information Science*, 33(2):115–133, 2007.

S. Fuchs. *The professional quest for truth: a social theory of science and knowledge*. SUNY series in science, technology, and society. State University of New York Press, Albany, 1992.

S. Fuchs. A Sociological Theory of Scientific Change. *Social Forces*, 71(4): 933–953, 1993a.

S. Fuchs. Three Sociological Epistemologies. *Sociological Perspectives*, 36(1): 23–44, 1993b.

E. Garfield. Is Information Retrieval in the Arts and Humanities Inherently Different from that in Science? The Effect that ISI®'s Citation Index for the Arts and Humanities is Expected to have on Future Scholarship. *The Library Quarterly*, 50(1):40–57, 1980.

E. Garfield, A. I. Pudovkin, and V. S. Istomin. Why do we need algorithmic historiography? *Journal of the American Society for Information Science and Technology*, 54(5):400–412, 2003.

W. Glänzel and U. Schoepflin. A bibliometric study of reference literature in the sciences and social sciences. *Information processing & management*, 35(1):31–44, 1999.

J. Gläser and G. Laudel. Governing Science. *European Journal of Sociology*, 57(01):117–168, 2016.

B. H. Good, Y.-A. de Montjoye, and A. Clauset. Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4), 2010.

A. Grafton. *The Footnote: a Curious History.* Harvard University Press, 1999.

M. Granovetter. The Strenght of Weak Ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

J. S. Grubb. When myths lose power: four decades of Venetian historiography. *The Journal of Modern History*, 58(1):43–94, 1986.

J. Guetzkow, M. Lamont, and G. Mallard. What is Originality in the Humanities and the Social Sciences? *American Sociological Review*, 69(2): 190–212, 2004.

C. Gumpenberger, J. Sorz, M. Wieland, and J. Gorraiz. Humanities and social sciences in the bibliometric spotlight – Research output analysis at the University of Vienna and considerations for increasing visibility. *Research Evaluation*, pages 1–8, 2016.

G. Halevi, H. Moed, and J. Bar-Ilan. Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation? review of the Literature. *Journal of Informetrics*, 11(3):823–834, 2017.

B. Hammarfelt. Interdisciplinarity and the intellectual base of literature studies: citation analysis of highly cited monographs. *Scientometrics*, 86 (3):705–725, 2011.

B. Hammarfelt. *Following the footnotes: a bibliometric analysis of citation patterns in literary studies.* PhD thesis, University of Uppsala, 2012a.

B. Hammarfelt. Harvesting footnotes in a rural field: citation patterns in Swedish literary studies. *Journal of Documentation*, 68(4):536–558, 2012b.

B. Hammarfelt. Using altmetrics for assessing research impact in the humanities. *Scientometrics*, 101(2):1419–1430, 2014.

B. Hammarfelt. Beyond Coverage: Toward a Bibliometrics for the Humanities. In M. Ochsner, S. E. Hug, and H.-D. Daniel, editors, *Research Assessment in the Humanities*, pages 115–131. Springer International Publishing, Cham, 2016.

B. Hammarfelt. Four Claims on Research Assessment and Metric Use in the Humanities. *Bulletin of the Association for Information Science and Technology*, 43(5):33–38, 2017.

A.-W. Harzing and S. Alakangas. Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, 106(2):787–804, 2016.

D. Heckmann, A. Frank, M. Arnold, P. Gietz, and C. Roth. Citation segmentation from sparse & noisy data: A joint inference approach with Markov logic networks. *Digital Scholarship in the Humanities*, 31(2):333–356, 2016.

R. Heinzkill. Characteristics of references in selected scholarly English literary journals. *The Library Quarterly*, 50(3):352–365, 1980.

R. Heinzkill. References in scholarly English and American literary journals thirty years later: A citation study. *College & Research Libraries*, 68(2): 141–154, 2007.

B. Hellqvist. Referencing in the humanities and its implications for citation analysis. *Journal of the American Society for Information Science and Technology*, 61(2):310–318, 2009.

D. Henriksen. The rise in co-authorship in the social sciences (1980–2013). *Scientometrics*, 107(2):455–476, 2016.

D. Hicks. The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics*, 44 (2):193–215, 1999.

D. Hicks. The four literatures of social science. In H. F. Moed, W. Glänzel, and U. Schmoch, editors, *Handbook of quantitative science and technology research*, pages 473–496. Springer, 2004.

T. Hitchcock. Confronting the Digital: Or How Academic History Writing Lost the Plot. *Cultural and Social History*, 10(1):9–23, 2013.

L. Horodowich. The new Venice: historians and historiography in the 21st century lagoon. *History Compass*, 2(1):1–27, 2004.

M.-h. Huang and Y.-w. Chang. Characteristics of research output in social sciences and humanities: From a research evaluation perspective. *Journal of the American Society for Information Science and Technology*, 59(11): 1819–1828, 2008.

S. E. Hug, O. Michael, and D. Hans-Dieter. A Framework to Explore and Develop Criteria for Assessing Research Quality in the Humanities. *International Journal for Education Law and Policy*, 10(1):55–68, 2014.

K. Hyland. Disciplinary Differences: Language Variation in Academic Discourses. In K. Hyland and M. Bondi, editors, *Academic Discourse Across Disciplines*, pages 17–45. Peter Lang, Bern, 2006.

J.-P. V. M. Hérubel. Citation Studies in the Humanities and Social Sciences: A Selective and Annotated Bibliography. *Collection Management*, 18(3-4): 89–137, 1994.

J.-P. V. M. Hérubel. Historical Bibliometrics: Its Purpose and Significance to the History of Disciplines. *Libraries & Culture*, 34(4):380–388, 1999.

J.-P. V. M. Hérubel and E. A. Goedeken. Using the *Arts and Humanities Citation Index* to Identify a Community of Interdisciplinary Historians: An Exploratory Bibliometric Study. *The Serials Librarian*, 41(1):85–98, 2001.

ICCU. Union Catalogue of Italian Libraries and Bibliographic Information - Reicat - GuidaSBN, 2016. URL `http://norme.iccu.sbn.it/index.php?title=Reicat&oldid=3034`. [Online; last accessed: January 9, 2017].

M. Infelise. Venezia e il suo passato. Storia, miti, 'fole'. In M. Isnenghi and S. Woolf, editors, *Storia di Venezia. L'Ottocento e il Novecento*, pages 967–988. Istituto dell'Enciclopedia Italiana Treccani, Rome, 2002.

M. Jacomy, T. Venturini, S. Heymann, and M. Bastian. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE*, 9(6):e98679, 2014.

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

C. Jones, M. Chapman, and P. C. Woods. The characteristics of the literature used by historians. *Journal of Librarianship and Information Science*, 4 (3):137–156, 1972.

S. N. Katz. Do Disciplines Matter? History and the Social Sciences. *Social Science Quarterly*, 76(4):863–877, 1995.

C. Kellsey and J. Knievel. Overlap between humanities faculty citation and library monograph collections, 2004–2009. *College & Research Libraries*, 73(6):569–583, 2012.

M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14(1):10–25, 1963.

Y.-M. Kim, P. Bellot, E. Faath, and M. Dacos. Automatic annotation of bibliographical references in digital humanities books, articles and blogs. In *Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing*, pages 41–48. ACM, 2011.

R. Klavans and K. W. Boyack. Toward a consensus map of science. *Journal of the American Society for Information Science and Technology*, 60(3): 455–476, 2009.

R. Klavans and K. W. Boyack. Which Type of Citation Analysis Generates the Most Accurate Taxonomy of Scientific and Technical Knowledge? *Journal of the Association for Information Science and Technology*, 68(4): 984–998, 2016.

J. E. Knievel and C. Kellsey. Citation Analysis for Collection Development: A Comparative Study of Eight Humanities Fields. *The Library Quarterly: Information, Community, Policy*, 75(2):142–168, 2005.

W. M. Kolasa. Specific character of citations in historiography (using the example of Polish history). *Scientometrics*, 90(3):905–923, 2012.

M. Körner, B. Ghavimi, P. Mayr, H. Hartmann, and S. Staab. Evaluating Reference String Extraction Using Line-Based Conditional Random Fields: A Case Study with German Language Publications. In M. Kirikova, K. Nørvåg, G. A. Papadopoulos, J. Gamper, R. Wrembel, J. Darmont, and S. Rizzi, editors, *New Trends in Databases and Information Systems*, volume 767, pages 137–145. Springer International Publishing, Cham, 2017.

H. Kreuzman. A co-citation analysis of representative authors in philosophy: Examining the relationship between epistemologists and philosophers of science. *Scientometrics*, 50(3):525–539, 2001.

F. Krämer. Shifting Demarcations: An Introduction. *History of Humanities*, 3(1):5–14, 2018.

T. S. Kuhn. *The structure of scientific revolutions*. University of Chicago Press, Chicago, 3rd edition, 1996.

E. Kulczycki, T. C. E. Engels, J. Pölönen, K. Bruun, M. Dušková, R. Guns, R. Nowotniak, M. Petr, G. Sivertsen, A. Istenič Starčič, and A. Zuccala. Publication patterns in the social sciences and humanities: evidence from eight European countries. *Scientometrics*, 2018.

S. Kyvik and I. Reymert. Research collaboration in groups and networks: differences across academic fields. *Scientometrics*, 113(2):951–967, 2017.

J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of ICML*, pages 282–289, 2001.

V. Larivière, É. Archambault, Y. Gingras, and É. Vignola-Gagné. The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities. *Journal of the American Society for Information Science and Technology*, 57(8):997–1004, 2006a.

V. Larivière, Y. Gingras, and É. Archambault. Canadian collaboration networks: A comparative analysis of the natural sciences, social sciences and the humanities. *Scientometrics*, 68(3):519–533, 2006b.

V. Larivière, É. Archambault, and Y. Gingras. Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004). *Journal of the American Society for Information Science and Technology*, 59(2):288–296, 2008.

V. Larivière, Y. Gingras, and E. Archambault. The decline in the concentration of citations, 1900-2007. *Journal of the American Society for Information Science and Technology*, 60(4):858–862, 2009.

L. Leydesdorff. The relations between qualitative theory and scientometric methods in science and technology studies: Introduction to the topical issue. *Scientometrics*, 15(5-6):333–347, 1989.

L. Leydesdorff and A. Salah. Maps on the basis of the Arts & Humanities Citation Index: The journals Leonardo and Art Journal versus "digital humanities" as a topic. *Journal of the Association for Information Science and Technology*, 61(4):787–801, 2010.

L. Leydesdorff, B. Hammarfelt, and A. Salah. The structure of the Arts & Humanities Citation Index: A mapping on the basis of aggregated citations among 1,157 journals. *Journal of the American Society for Information Science and Technology*, 62(12):2414–2426, 2011.

C.-S. Lin, Y.-F. Chen, and C.-Y. Chang. Citation functions in social sciences and humanities: Preliminary results from a citation context analysis of Taiwan's history research journals. *Proceedings of the American Society for Information Science and Technology*, 50(1):1–5, 2013.

Y. LindholmRomantschuk and J. Warner. The role of monographs in scholarly communication: an empirical study of philosophy, sociology and economics. *Journal of Documentation*, 52(4):389–404, 1996.

A. J. M. Linmans. Why with bibliometrics the Humanities does not need to be the weakest link: Indicators for research evaluation based on citations, library holdings, and productivity measures. *Scientometrics*, 83(2):337–354, 2009.

M. Liu, X. Hu, and J. Li. Knowledge flow in China's humanities and social sciences. *Quality & Quantity*, 52(2):607–626, 2017.

P. Lopez. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Research and Advanced Technology for Digital Libraries*, pages 473–474. Springer, 2009.

M. S. Lowe. Reference analysis of the American Historical Review. *Collection Building*, 22(1):13–20, 2003.

T. Luukkonen. Why has Latour's theory of citations been ignored by the bibliometric community? Discussion of sociological interpretations of citation analysis. *Scientometrics*, 38(1):27–37, 1997.

M. D. Marchi and E. Lorenzetti. Measuring the impact of scholarly journals in the humanities field. *Scientometrics*, 106(1):253–261, 2015.

I. Marshakova Shaikevich. System of Document Connections Based on References. *Scientific and Technical Information Serial of VINITI*, 6(2):3–8, 1973.

K. W. McCain. Citation Patterns in the History of Technology. *Library & Information Science Research*, 9:41–59, 1987.

W. McCarty. Humanities computing. *Encyclopedia of library and information science*, pages 1224–1236, 2003.

W. McCarty. Becoming Interdisciplinary. In S. Schreibman, R. Siemens, and J. Unsworth, editors, *A New Companion to Digital Humanities*, pages 67–83. John Wiley & Sons, Ltd, Chichester, 2015.

R. K. Merton. The Matthew Effect in Science. *Science*, 159(3810):56–63, 1968.

R. K. Merton. *The sociology of science: theoretical and empirical investigations*. University of Chicago Press, Chicago, 4th edition, 1974.

J. Mingers and L. Leydesdorff. A review of theory and practice in scientometrics. *European Journal of Operational Research*, 246(1):1–19, 2015.

H. F. Moed, W. J. M. Burger, J. G. Frankfort, and A. F. J. Van Raan. The use of bibliometric data for the measurement of university research performance. *Research policy*, 14(3):131–149, 1985.

P. Mongeon and A. Paul-Hus. The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, 106(1):213–228, 2016.

S. A. Morris and B. Van der Veer Martens. Mapping research specialties. *Annual review of information science and technology*, 42(1):213–295, 2008.

A. Nederhof, J. Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A review. *Scientometrics*, 66(1):81–100, 2006.

M. E. J. Newman. *Networks: an introduction*. Oxford University Press, Oxford, 2010.

M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2), 2004.

D. S. Nolen and H. A. Richardson. The Search for Landmark Works in English Literary Studies: A Citation Analysis. *The Journal of Academic Librarianship*, 42(4):453–458, 2016.

M. Ochsner, S. E. Hug, and H.-D. Daniel. Humanities Scholars' Conceptions of Research Quality. In M. Ochsner, S. E. Hug, and H.-D. Daniel, editors, *Research Assessment in the Humanities*, pages 43–69. Springer International Publishing, Cham, 2016a.

M. Ochsner, S. E. Hug, and H.-D. Daniel, editors. *Research Assessment in the Humanities*. Springer International Publishing, Cham, 2016b.

N. Okazaki. CRFsuite: a fast implementation of Conditional Random Fields (CRFs), 2007. URL `www.chokkan.org/software/crfsuite`.

M. Olensky, M. Schmidt, and N. J. van Eck. Evaluation of the citation matching algorithms of CWTS and iFQ in comparison to the Web of Science. *Journal of the Association for Information Science and Technology*, 67(10):2550–2564, 2016.

A. Perianes-Rodriguez, L. Waltman, and N. J. van Eck. Constructing bibliometric networks: A comparison between full and fractional counting. *Journal of Informetrics*, 10(4):1178–1195, 2016.

O. Persson. The intellectual base and research fronts of "jasis" 1986-1990. *Journal of the American Society for Information Science*, 45(1):31, 1994.

H. Piwowar, J. Priem, V. Larivière, J. P. Alperin, L. Matthias, B. Norlander, A. Farley, J. West, and S. Haustein. The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ*, 6:e4375, 2018.

J. H. Plumb, editor. *Crisis in the humanities*. Penguin, London, 1964.

M. Polanyi. *Personal knowledge: towards a post-critical philosophy*. Routledge, London, 1997.

C. Povolo. The Creation of Venetian Historiography. In J. J. Martin and D. Romano, editors, *Venice Reconsidered: The History and Civilization of* *an Italian City-State, 1297-1797*, pages 491–519. Johns Hopkins University Press, Baltimore, 2002.

A. Prasad, M. Kaur, and M.-Y. Kan. Neural parscit: A deep learning based reference string parser. *Forthcoming in International Journal on Digital Libraries*, 2018.

E. Reale, D. Avramov, K. Canhial, C. Donovan, R. Flecha, P. Holm, C. Larkin, B. Lepori, J. Mosoni-Fried, E. Oliver, E. Primeri, L. Puigvert, A. Scharnhorst, A. Schubert, M. Soler, S. Soòs, T. Sordé, C. Travis, and R. Van Horik. A review of literature on evaluating the scientific, social and political impact of social sciences and humanities research. *Research Evaluation*, 2017.

J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1), 2006.

D. Rodrigues Alves, G. Colavizza, and F. Kaplan. Deep Reference Mining from Scholarly Literature in the Arts and Humanities. *Forthcoming in Frontiers in Research Metrics & Analytics*, 2018.

M. Romanello. *From Index Locorum to Citation Network: an Approach to the Automatic Extraction of Canonical References and its Applications to the Study of Classical Texts.* PhD thesis, King's College London, 2015.

M. Romanello. Exploring Citation Networks to Study Intertextuality in Classics. *Digital Humanities Quarterly*, 10(2):1–12, 2016.

M. Romanello and G. Colavizza. dhlab-epfl/LinkedBooksMonographs: LinkedBooksMonographs (version 1.1), feb 2017. URL `https://doi.org/10.5281/zenodo.377047`.

S. Schreibman, R. G. Siemens, and J. Unsworth, editors. *A companion to digital humanities.* Number 26. Blackwell Pub, Malden, 2004.

Sci2 Team. Science of Science (Sci2) Tool. Indiana University and SciTech Strategies, 2009. URL `https://sci2.cns.iu.edu`.

H. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265–269, 1973.

C. P. Snow. *The Two Cultures and the Scientific Revolution.* Cambridge University Press, London, 1959.

K. Sparck Jones, S. Walker, and S. Robertson. A probabilistic model of information retrieval: development and comparative experiments. part 1. *Information Processing & Management*, 36(6):779–808, 2000a.

K. Sparck Jones, S. Walker, and S. Robertson. A probabilistic model of information retrieval: development and comparative experiments. part 2. *Information Processing & Management*, 36(6):809–840, 2000b.

P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. BRAT: A Web-based Tool for NLP-assisted Text Annotation. In *Proceedings of the Demonstrations at the* 13^{th} *Conference of the European Chapter of the Association for Computational Linguistics*, EACL, pages 102–107. Association for Computational Linguistics, 2012.

S. Stone. Humanities Scholars: Information Needs and Uses. *Journal of Documentation*, 38(4):292–313, 1982.

N. W. Storer. The hard sciences and the soft: some sociological observations. *Nulletin of the Medical Library Association*, 55(1):75–84, 1967.

C. R. Sugimoto and B. Cronin, editors. *Theories of informetrics and scholarly communication: a Festschrift in honor of Blaise Cronin.* De Gruyter, Berlin, 2016.

C. R. Sugimoto and S. Weingart. The kaleidoscope of disciplinarity. *Journal of Documentation*, 71(4):775–794, 2015.

C. A. Sula and M. Miller. Citations, contexts, and humanistic discourse: Toward automatic extraction and classification. *Literary and Linguistic Computing*, 29(3):452–464, 2014.

I. Tahamtan and L. Bornmann. Core elements in the process of citing publications: Conceptual overview of the literature. *Journal of Informetrics*, 12(1):203–216, 2018.

S. Talja and H. Maula. Reasons for the use and non-use of electronic journals and databases: A domain analytic study in four scholarly disciplines. *Journal of Documentation*, 59(6):673–691, 2003.

M. Thelwall. Do Mendeley reader counts indicate the value of arts and humanities research? *Journal of Librarianship and Information Science*, pages 1–8, 2017.

M. Thelwall and M. M. Delgado. Arts and humanities research evaluation: no metrics please just data. *Journal of Documentation*, 71(4):817–833, 2015.

J. W. Thompson. The death of the scholarly monograph in the humanities? Citation patterns in literary scholarship. *Libri*, 52(3):121–136, 2002.

D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, and L. Bolikowski. CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(4):317–335, 2015.

D. Tkaczyk, A. Collins, P. Sheridan, and J. Beel. Evaluation and comparison of open source bibliographic reference parsers: A business use case. *arXiv preprint arXiv:1802.01168*, 2018.

C. B. Trace and U. P. Karadkar. Information management in the humanities: Scholarly processes, tools, and the construction of personal collections. *Journal of the Association for Information Science and Technology*, 68(2): 491–507, 2017.

P. Trowler. Depicting and researching disciplines: strong and moderate essentialist approaches. *Studies in Higher Education*, 39(10):1720–1731, 2013.

P. Trowler. Academic tribes and territories: The theoretical trajectory. *Österreichische Zeitschrift Für Geschichtswissenschaften*, 25(3):17–26, 2014.

C.-C. Tsai, E. A. Corley, and B. Bozeman. Collaboration experiences across scientific disciplines and cohorts. *Scientometrics*, 108(2):505–529, 2016.

J. Turner. *Philology The Forgotten Origins of the Modern Humanities.* Princeton University Press, Princeton, NJ, 2015.

I. R. Tyrrell. *Historians in public: the practice of American history, 1890-1970.* University of Chicago Press, Chicago, 2005.

T. van Leeuwen. The application of bibliometric analyses in the evaluation of social science research. Who benefits from it, and why it is still feasible. *Scientometrics*, 66(1):133–154, 2006.

F. T. Verleysen and T. L. B. Ossenblok. Profiles of monograph authors in the social sciences and humanities: an analysis of productivity, career stage, co-authorship, disciplinary affiliation and gender, based on a regional bibliographic database. *Scientometrics*, 111(3):1673–1686, 2017.

F. T. Verleysen and A. Weeren. Mapping Diversity of Publication Patterns in the Social Sciences and Humanities: An Approach Making Use of Fuzzy Cluster Analysis. *Journal of Data and Information Science*, 1(4):33–59, 2016a.

F. T. Verleysen and A. Weeren. Clustering by publication patterns of senior authors in the social sciences and humanities. *Journal of Informetrics*, 10 (1):254–272, 2016b.

L. Waltman. A review of the literature on citation impact indicators. *Journal of Informetrics*, 10(2):365–391, 2016.

R. Watson-Boone. The information needs and habits of humanities scholars. *Reference Quarterly*, 34(2):203–215, 1994.

S. B. Weingart. Finding the History and Philosophy of Science. *Erkenntnis*, 80(1):201–213, 2015.

C. Wellmon. *Organizing Enlightenment: information overload and the invention of the modern research university.* Johns Hopkins University Press, Baltimore, 2015.

R. Whitley. *The intellectual and social organization of the sciences.* Oxford University Press, Oxford, 1984.

R. Whitley. *The intellectual and social organization of the sciences.* Oxford University Press, Oxford ; New York, 2nd edition, 2000.

S. E. Wiberley Jr. Humanities Literatures and Their Users. In *Encyclopedia of Library and Information Sciences*, pages 2197–2204, 2010.

P. Williams, I. Stevenson, D. Nicholas, A. Watkinson, and I. Rowlands. The role and future of the monograph in arts and humanities research. *Aslib Proceedings*, 61(1):67–82, 2009.

W. Windelband. *Geschichte und Naturwissenschaft.* Heitz, Strassburg, 1904.

P. Wouters. *The citation culture.* PhD thesis, University of Amsterdam, 1999.

P. Wouters and L. Leydesdorff. Has Price's dream come true: Is scientometrics a hard science? *Scientometrics*, 31(2):193–222, 1994.

J. Wu, K. Williams, H.-H. Chen, M. Khabsa, C. Caragea, A. Ororbia, D. Jordan, and C. L. Giles. Citeseerx: Ai in a digital library search engine. In *Innovative Applications of AI Conference*, 2014.

S. Wyatt, S. Milojevic, H. Park, and L. Leydesdorff. *The Intellectual and Practical Contributions of Scientometrics to STS*, pages 87–112. The MIT Press, 4th edition, 2016.

C. Zerby. *The devil's details: a history of footnotes.* Touchstone Book, New York, 2003.

J. M. Ziman. *Public knowledge: an essay concerning the social dimension of science.* Cambridge University Press, Cambridge, 1968.

G. Zordan. *Repertorio di storiografia veneziana: testi e studi.* Il Poligrafo, Padova, 1998.

A. Zuccala, R. Guns, R. Cornacchia, and R. Bod. Can we rank scholarly book publishers? A bibliometric experiment with the field of history. *Journal of the Association for Information Science and Technology*, 66(7):1333–1347, 2015.

H. Zuckern and R. H. Merton. Age Aging and Age Structure in Science. In *The Sociology of Science. Theoretical and Empirical Investigations*, pages 497–539. University of Chicago Press, Chicago, 1973.

L. Šubelj, N. J. van Eck, and L. Waltman. Clustering Scientific Publications Based on Citation Relations: A Systematic Comparison of Different Methods. *PLOS ONE*, 11(4):e0154404, 2016.