Mapping Cultural Connectivity: Network and Embeddings Analysis Applied to English Wikipedia's Internal Link Structure

PhD thesis by

Paschalis Agapitos

Universidad del País Vasco / Euskal Herriko Unibertsitatea

Supervised by

Gustavo Ariel Schwartz and Juan Luis Suárez

2026

Table of Contents

Table of Contents

Table of Contents

List of Figures

List of Figures

Abstract

Cultural relationships between entities often leave no direct trace in the historical record, surfacing instead as diffuse patterns of association across large bodies of data. This thesis develops a computational approach, combining network analysis and distributional semantics with humanistic interpretation, to recover such patterns from Wikipedia's link network and examine how they are encoded within large-scale cultural data.

Chapter 1 establishes the conceptual and methodological foundations of the thesis. It opens by framing the core problem: many culturally meaningful relationships are not recorded as direct links between two entities but emerge indirectly, as patterns of association across large bodies of data that traditional close reading was never designed to handle. The chapter then introduces how this thesis addresses this problem by using "focus reading" that combines computational methods, network analysis and distributional semantics, to detect patterns at scale with humanistic frameworks to interpret their significance, situating this approach within digital humanities and cultural data analytics. It explains why the network is treated as a deliberate epistemological choice rather than a neutral mirror of reality, and defines what distinguishes a specifically cultural network: one in which relationships between people are mediated by shared objects, ideas, and institutions. The chapter argues that Wikipedia's network constitutes a suitable, if normatively filtered, instance of such a network, and closes by motivating the study and presenting the three research questions that structure the remaining chapters.

Chapter 2 presents the methodology shared across the three studies and the specific techniques each requires. It explains that all three draw on the same analytical toolkit: network science, which examines the structure of systems of connected entities, and distributional semantics, which represents meaning through context, with each study applying these tools at a different scale. For the seventeenth-century study, the chapter

introduces a measure of structural relatedness inspired by the Normalised Google Distance, which infers conceptual closeness between articles from shared linking patterns rather than from direct links alone, and pairs this measure with bootstrapping to test whether observed differences are statistically robust. It then introduces WikiTextGraph, the open-source tool developed for the data mining phase of this thesis, describing how it parses Wikipedia dumps efficiently and reconstructs the internal link network in a way that is reproducible and accessible to non-specialists. Finally, it outlines the procedure for the large-scale biographical study, including the use of Wikidata to reliably detect nearly two million biographies and the combination of network topology with semantic embeddings to recover long-range patterns across twenty-five centuries.

Chapter 3 presents the results, divided into two parts corresponding to the two studies conducted.

Section 3.1 reports the results of the first study, which serves as a proof of concept for the overall approach. It asks whether wikilink-based network analysis can recover historically meaningful patterns within a defined period, testing this on the cultural network connecting Art, Science, and Philosophy in seventeenth-century Europe. Rather than analysing a single triad of figures, as earlier studies had done, it averages across 465 triads sampled from the period, so that the results describe the period itself rather than any one configuration of individuals. The analysis operates at three scales. At the level of the whole network, high modularity shows that the period divides cleanly into three disciplinary clusters. At the cluster level, the measures reveal that Philosophy was the most internally cohesive discipline and that the connection between Science and Philosophy was by far the strongest, while Art occupied a peripheral position, findings that align with what historians have independently established about the period. At the level of individual nodes, the study identifies two distinct statistical patterns of connectivity, a core-periphery structure within disciplines and a heavy-tailed distribution across them and recovers central figures and concepts

that were never chosen as starting points. Together, these results confirm that Wikipedia's network encodes recoverable cultural structure and that the method is sensitive to genuine historical change.

Section 3.2 reports the large-scale study, which extends the approach from a single period to roughly 1.9 million biographies spanning twenty-five centuries, combining network analysis with semantic embeddings. The network analysis shows that biographies mostly link to others born within the same or adjacent century, a pattern of temporal homophily so strong that the average gap between linked individuals, about thirty-four years, closely matches the span of a human generation. The sparse long-range links that remain are not random: connections from later scientists back to Classical figures trace the documented transmission of ancient knowledge, while the choice of how links are counted determines whether the network tells a story of historical continuity or of rupture, a reminder that such methodological choices are interpretive rather than neutral. The semantic analysis then finds that centuries close in time occupy nearby regions of the embedding space, conventional historical periods re-emerge from the data without being imposed, and semantic similarity declines gradually with temporal distance, a regularity termed here the temporal Tobler pattern. The outliers to that trend surface periods that are distant in time yet conceptually close, such as the link between thirteenth-century science and Classical antiquity through the Arabic-Latin translation movement. Taken together, the section shows that both the network and the semantic structure of Wikipedia's biographical network are organised by time, and that the patterns recoverable from that network reflect not only historical processes but the representational choices (methodological and editorial) through which those processes have been made legible at scale.

Chapter 4 draws the three studies together and reflects on what they collectively establish. It revisits each research question in turn: the first demonstrated that Wikipedia's link network retains historically meaningful structure at the level of a period, the second built the computational infrastructure needed to work at scale, and

the third combined both to trace temporal and semantic patterns across nearly two million biographies. From this, the chapter draws out the theoretical contributions, chiefly the treatment of a historical period as a measurable analytical object through triad averaging, and the demonstration that the structural and semantic dimensions of a cultural network are complementary rather than redundant. It then turns, with care, to the limitations, acknowledging that Wikipedia is a socially situated representation shaped by gender, geographic, linguistic, and notability biases, so that the patterns recovered describe how cultural memory has been encoded rather than the past itself. The chapter closes by setting out directions for future work, including identifying core figures within disciplines, modelling the network's generative mechanisms, and extending the framework to other periods, languages, and data sources.

This thesis was carried out by the author at the Centro de Física de Materiales (CFM) under the supervision of Dr. Gustavo Ariel Schwartz (CFM) and Dr. Juan Luis Suárez (CulturePlex Lab, London, Ontario, Canada), with funding provided by the Donostia International Physics Centre (DIPC) through the Mestizajes project.

This thesis is the result of three studies:

1. Miccio LA, **Agapitos P**, Gamez-Perez C, et al (2025) Wikipedia as a cultural lens: a quantitative approach for exploring cultural networks. Humanit Soc Sci Commun 12:462. https://doi.org/10.1057/s41599-025-04772-5
2. **Agapitos P**, Suárez J-L, Schwartz GA (2025) WikiTextGraph: A Python Tool for Parsing Multilingual Wikipedia Text and Graph Extraction. Journal of Open Research Software 13(1):17. https://doi.org/10.5334/jors.572
3. **Agapitos P**, Suárez J-L, Schwartz GA (2026) Temporal and Semantic Patterns in Wikipedia's Network of Biographies. (to be submitted)

Objective

Cultural relationships, understood here as the connections between people, ideas, and artefacts, are complex by nature. These relationships can show up in countless forms, from religious practice to everyday interaction. This thesis focuses on the ones captured in structured online knowledge repositories. Within these repositories, these connections are often distributed across large volumes of data and are frequently not directly observable. For instance, many such relationships are not recorded as explicit connections between two elements. Instead, they emerge from patterns of association across interconnected collections of data. Finding those patterns requires tools that can work at a scale and complexity that traditional humanities methods were not built for, since those methods were designed for close reading of small corpora and not for the volumes of cultural data that is now available. This gap between the scale of available data and the scope of existing analytical approaches motivates the present thesis.

Wikipedia's internal link network provides a particularly suitable basis for addressing this challenge. As one of the largest collaboratively built knowledge repositories, Wikipedia encodes relationships between cultural entities[1] through millions of wikilinks.[2] These wikilinks reflect collective editorial decisions made over time in a process that does not have an established endpoint. Because of this collaborative and cumulative process, the wikilink network captures not only explicit factual connections

[1] A cultural entity is any discrete topic that can be the subject of a Wikipedia article, such as a person, institution, event, or place. It is not a direct representation of its real-world referent; rather, it is an abstract knowledge entity embedded in a network of relations, comprising multiple layers like social, scientific, and political contexts, along with connections to other cultural entities. It emerges from two sources of knowledge: the explicit knowledge produced by experts, and the implicit knowledge derived from quantitatively analysing Wikipedia's wikilinks as a network [1].

[2] A *wikilink* is an internal (hyperlink) in a wiki that points to another page or section within the same wiki version.

but also implicit patterns of association that reflect how contributors collectively organise knowledge.

The aim of this thesis is to develop and apply methods, both new and existing, to identify, quantify, and interpret connectivity patterns between cultural entities using Wikipedia's network as the main data source.[3] This aim is pursued through three successive contributions, each building on the previous one. The first examines whether wikilink-based network analysis can recover meaningful connectivity patterns between cultural entities within a defined historical period. More specifically, it investigates the relationships between Art, Science, and Philosophy in seventeenth-century Europe. This serves as a methodological proof of concept by testing whether the proposed analytical framework can produce historically meaningful results within a well-documented domain. The second addresses a prerequisite for scaling the analysis. It introduces WikiTextGraph, an open-source software tool for extracting text and wikilink networks from multilingual Wikipedia dumps. WikiTextGraph is built for reproducibility and computational efficiency, and it provides the infrastructure needed to extend the analysis beyond a single historical period. The third contribution applies the combined previous steps to suggest a framework for a large-scale analysis of Wikipedia's biographical subnetwork. Through this framework it examines the temporal and semantic patterns that emerge across approximately two million biographies spanning twenty-five centuries.

[3] Throughout this thesis, all references to "Wikipedia's network" denote the wikilink network of the English-language edition of Wikipedia.

Objective

1. Introduction

> "By thinking about culture as data we open up opportunities both for new analytical processes and for new areas of discourse and engagement."
>
> Ahnert et al., *The Network Turn*, p. 56 [2]

Over the past two decades, the growing availability of large-scale digital data has substantially expanded how we study culture. Such data arise from a range of sources (e.g., online platforms, digitised archives, and collaborative knowledge systems) and it captures interactions between people, ideas, and artefacts at a scale that was simply not possible before [3, 4]. As a result, cultural information can now be understood as a vast, interconnected system.

However, this expansion in scale introduces a methodological challenge. Many relationships between cultural entities are not recorded explicitly. Instead, they emerge from patterns of co-occurrence and association across large, interconnected datasets [3, 5–8]. Such relationships are referred to here as *indirect*: they are present in the data but not directly observable. They have to be inferred from the structure of the data as a whole. To make this distinction clearer, consider two types of relationship. An explicit relationship takes the form: *entity A is directly connected to entity B*. An indirect relationship, by contrast, might be inferred from the observation that *entity A and entity B are each independently connected to entity C*. In the latter case, the connection between A and B is not recorded anywhere directly; it only becomes visible when the broader network of associations is considered. This means that recovering the full relational structure of a cultural dataset requires attending not only to direct connections but also to indirect ones — relationships that exist through intermediary elements.

Both tasks (identifying explicit relationships and inferring indirect ones from patterns of co-occurrence) become increasingly difficult as the complexity of interactions among a large number of cultural entities grows [8]. Moreover, because the number of possible pairwise interactions between cultural entities grows combinatorially with the number of elements themselves, the higher-order structures that organise cultural knowledge become progressively harder to recover from the data. Methods developed within the humanities for close, interpretive reading of small corpora are not designed to operate at this level of scale and complexity — not because they lack analytical rigour, but because they were built for a fundamentally different mode of engagement with cultural material. A different set of analytical tools is therefore required.

However, the same expansion in digital data that generates this methodological challenge also creates new conditions for addressing it. Publicly accessible corpora and digital platforms provide access not only to cultural artefacts themselves, but also to the relational structures connecting them; structures that would be difficult to recover at this scale through manual analysis alone. When combined with computational methods capable of processing large, interconnected datasets, this relational data makes it possible to detect, describe, and compare patterns of connections between cultural entities that would otherwise remain inaccessible [9].

This convergence of data availability and analytical capacity has produced methodological approaches that apply quantitative tools to questions traditionally framed within the humanities [4]. Rather than treating these as competing orientations, this thesis treats quantitative rigour and humanistic interpretation as complementary. The former offers systematic pattern detection and quantification at scale. The latter provides the conceptual frameworks needed to interpret what those patterns mean. In practice, this means using computational methods to detect and quantify patterns of connection across large collections of cultural entities and then drawing on humanistic frameworks to interpret what those patterns reveal about the organisation of cultural knowledge.

1.1. Close and Distant Reading

The tension between interpreting individual works and analysing large collections of cultural entities is not new to digital methods. Close reading treats meaning as something located in the specific features of individual texts. It involves careful, recursive interpretation of a small number of works, with attention to language, syntax, and internal structure [10–15]. Developed within 20th-century Anglo-American literary criticism, the approach treats each text as a self-contained object whose meaning emerges from its formal organisation. Scale and particularity are its defining commitments: close reading assumes that meaning cannot be recovered without sustained, repeated engagement with the individual work.

Distant reading looks for meaning in patterns that emerge across many texts. Franco Moretti coined the term around 2000 to describe a mode of literary analysis that operates at a fundamentally different scale [16, 17]. Instead of explaining individual works, it aims to reveal large-scale structures: the rise and fall of genres, the spread of literary forms across languages, and the long-term dynamics of what Moretti called the literary world-system. The individual text is set aside, because the patterns of interest only become visible when many texts are examined together.

Although first framed as opposing alternatives, the two approaches are now widely treated as complementary. Matthew Jockers, for example, uses computational analysis of large corpora to generate hypotheses about literary history, then refines those hypotheses through close engagement with individual texts [18]. Ted Underwood places this combination within a longer tradition of book history and literary sociology that predates digital humanities as a distinct field [19, 20]. Together, these arguments show that aggregate and interpretive methods are not rivals but different levels of analysis that can strengthen one another.

Yet treating the two as simply complementary does not fully resolve the tension between them. Each approach retains structural limitations that the other cannot

compensate for. For example, close reading, by definition, cannot operate at scale and therefore fails to detect large-scale structural patterns; distant reading, conversely, can identify such patterns but lacks the interpretive apparatus to ground those patterns in specific cultural and historical meaning. What is needed is a method that operates at both levels simultaneously: one that uses large-scale pattern detection to identify what is worth examining closely and then deploys humanistic interpretation to make those patterns meaningful. This is the logic that motivates focus reading [21].

In practice, focus reading works like this. Rather than reading everything closely, which is impossible at scale, or reading nothing closely, which leaves patterns without interpretive content, the analyst first uses network analysis to scan a large corpus and identify its most significant nodes. These nodes represent the people, ideas, and works most densely connected to the subject under investigation: the points through which the greatest volume of intellectual traffic passes. The analyst then reads those specific nodes closely, drawing on established humanistic knowledge to interpret what the network has surfaced. The network identifies where to look. Humanistic interpretation determines what it means.

This thesis adopts focus reading as its primary analytical orientation. Its main object is not any single Wikipedia article but the structural patterns that emerge across millions of them, examined through complex network analysis. That focus matters methodologically. Wikipedia articles are not authored texts in the conventional sense: their meaning is produced through collective editorial negotiation rather than through the stylistic choices of a single author. Formal literary close reading does not apply directly. What the thesis retains is the interpretive commitment that underlies close reading, namely the expectation that aggregate patterns only become analytically meaningful when read through established humanistic frameworks. Over the past two decades, this combination of large-scale pattern detection and humanistic interpretation has been formalised into two related disciplinary fields.

1.2. Digital Humanities & Cultural Data Analytics

Two related disciplinary traditions formalise the complementarity between aggregate pattern detection and humanistic interpretation established in *section 1.1*, and together provide the intellectual context for the approach developed in this thesis: digital humanities and cultural analytics. While both draw on quantitative methods, they differ in scope and emphasis in ways that are consequential for the analysis developed here.

Digital humanities (hereafter DH) denotes the application of computational methods to research questions across the humanities including history, literary studies, linguistics, philosophy, and cultural studies [22–24]. In practice, this encompasses techniques such as text mining, natural language processing, network analysis, and statistical modelling, applied to textual, visual, and other cultural materials at a scale that exceeds what manual analysis can address. The field is defined less by any single method than by a shared orientation: computational tools are developed and used to identify structures and patterns in large collections of data, while research questions, interpretive categories, and evaluative frameworks remain grounded in humanistic inquiry. This dual commitment is what distinguishes DH from purely technical approaches to data analysis and from traditional humanities scholarship alike and it is within this tradition that this thesis operates.

Within this broader landscape, *cultural analytics* (hereafter CA) represents a more specifically focused area of inquiry. Where DH encompasses the full range of humanistic disciplines and methods, CA is concerned primarily with the quantitative analysis and visualisation of large cultural datasets, with the aim of identifying patterns in cultural production, circulation, and change over time [4, 6, 25]. The materials that constitute cultural data in this context are diverse, ranging from digitised books, films, and photographs to natively digital content such as social media, video games, and collaborative online platforms [4]. However, a defining characteristic of the field is not

merely the breadth of its materials but the analytical implications of their scale: the volume of cultural data now available exceeds what any researcher can examine item by item. At this scale, computational methods are not merely convenient but analytically necessary, because they make it possible to pose questions about cultural patterns that close, item-level reading could not address.

However, this reorientation is not purely methodological. It reflects a broader epistemological shift in what constitutes the legitimate object of cultural inquiry: a move from the intensive study of a small number of exemplary works towards the systematic analysis of large populations of artefacts, practices, and representations [26]. Rather than treating individual texts or objects as windows onto culture, this perspective treats patterns across collections as analytically meaningful in their own right. The question is no longer only what a given artefact means, but what structural regularities emerge when many artefacts are examined together. Taken together, these two traditions establish the intellectual coordinates within which the present thesis is situated: a commitment to using computational methods — here, specifically network analysis and distributional semantics — to detect and quantify patterns of connectivity between cultural entities at scale, interpreted through humanistic frameworks concerned with how knowledge is structured and organised.

1.3. The Network as an Epistemological Choice

> "The network framework shapes how we interpret the world around us. Nothing is naturally a network; rather, networks are an abstraction into which we squeeze the world."
>
> Ahnert et al. *The Network Turn* [27]

The decision to represent cultural data as a network is not only driven by computational convenience. A network makes the relation between entities, rather than the individual entity itself, the unit of analysis, making it possible to see which

entities are constituted by, or derive their significance from, their relationships. This aspect is very important because, as mentioned earlier, many culturally meaningful relationships are indirect; they do not appear as a labelled connection between two entities, but they emerge from patterns of co-occurrence and aggregation across a large, interconnected dataset. A network representation is therefore the most appropriate way to study such structures, since it is designed precisely to make this kind of relational complexity tractable.

To understand why network representations are analytically valuable, it helps to consider what simpler alternatives cannot do. A basic list of cultural entities and their types (such as people, objects, or ideas) do nothing beyond naming those entities since it reveals nothing about how they relate to one another. Adding one more layer, the frequency distribution of types of cultural entities, captures only how often entities appear but conceals which ones cluster together or bridge disconnected domains (i.e., how they relate to each other). By contrast, a network makes the relational organisation of the data itself the object of inquiry without presupposing any particular form for the representation. As Ahnert et al [27] argue, network analysis makes it possible to measure relationships between many entities simultaneously and in multiple ways, enabling a form of engagement with complex systems that neither close reading nor statistical aggregation alone can achieve. As noted earlier, this thesis does not reject close reading or statistical analysis. Both are incorporated into its analytical pipeline. In this context, networks function as an analytical filter: a way of surfacing patterns latent in large volumes of interconnected data while simultaneously identifying which entities and relationships warrant closer humanistic scrutiny.

The filter metaphor also implies a shift in analytical scale, from individual nodes and their immediate relationships to the system as a whole. This is an explanatory gap that network analysis is well positioned to bridge. Some of the cultural patterns this thesis investigates, such as the clustering of intellectual domains, the structural positions of bridging figures, and shifting connectivity patterns between fields over time, are

neither visible at the level of any single entity nor recoverable from aggregate statistics that ignore relational structure. They are properties of the network itself. The capacity to move between part and whole is therefore the central epistemological contribution the network form provides, and the reason the representational choice made here is analytically motivated rather than merely convenient.

Establishing that a network representation is appropriate for interconnected cultural phenomena does not, however, mean that every network qualifies as a cultural network. This distinction matters because network analysis, as a general method, is agnostic about the nature of the entities and relationships it models. It can equally represent traffic flows, protein interactions, and social media followings. If the claim is that the approach pursued here is suited to cultural inquiry specifically, then a principled definition of what constitutes a cultural network is required.

For the purposes of this thesis, the most useful definition is offered by Suárez et al. [28], who define a cultural network as a multimodal network in which at least two node types represent persons and cultural objects respectively. In such networks, relationships between persons are mediated by cultural objects (ideas, artefacts, institutions) rather than by direct social ties alone. This definition does several important pieces of work simultaneously, and it is worth unpacking each of them in turn.

First, the requirement of multimodality rules out networks composed of a single node type. A network of persons linked only by co-authorship captures social proximity but not the cultural content through which that proximity is constituted. What makes a network *cultural* is precisely that the objects circulating between persons, whether texts, concepts, artworks, or institutions, are present as nodes, not merely as labels attached to edges. Without this, cultural relationships collapse into purely social ones, and the patterns this thesis seeks to examine are obscured.

Second, requiring that relationships between persons be mediated by cultural objects shifts the unit of analysis away from individual actors and towards the shared cultural infrastructure through which they are connected. Two scholars may never have corresponded directly, yet both may link extensively to the same set of concepts, fields, and institutions. In a cultural network, that shared referential context is itself a relationship, and one of a specific kind: it reveals something about the organisation of knowledge rather than merely the topology of acquaintance.

Third, by requiring that cultural objects include not only artefacts but also ideas and institutions, the definition accommodates the kind of heterogeneous, semantically rich data that characterises encyclopaedic sources such as Encyclopaedia Britannica and Wikipedia. A network spanning only persons and texts would miss the institutional and conceptual infrastructure through which cultural knowledge is organised and transmitted. A network spanning persons, ideas, institutions, and events simultaneously allows the analyst to examine how different types of cultural entity relate to one another, and whether those patterns shift across historical time or disciplinary domain.

Not every large-scale network assembled from cultural data will therefore qualify as a cultural network in this sense. The defining criterion is not size, nor even the presence of humanistically recognisable entities, but the mediated structure of relationships: persons connected through objects, fields intersecting through shared concepts, institutions rendered visible through the figures and ideas they concentrate. When this structure is present, the network does more than organise data efficiently. It encodes the relational logic through which a community has distributed and structured its collective knowledge.

1.4. Wikipedia's network as a cultural network

The methodological framework established in *section 1.2* (i.e., combining computational pattern detection with humanistic interpretation within the general

tradition of DH and CA) places specific demands on an empirical object: it must be large enough to require computational methods, structured enough to support systematic and reproducible analysis, and it must be a cultural network. English Wikipedia meets all three conditions. With almost 7.2 million[4] articles, it exceeds the scale where manual analysis is practical. Its articles are organised according to stable editorial conventions and interconnected through a consistent hyperlinking system. And because it is produced through ongoing collective editorial negotiation among a large and diverse community of contributors, its structure is not an engineering artefact but a direct expression of how that community organises and relates cultural knowledge [30]. This is the reason why Wikipedia's network constitutes the primary empirical object of this thesis.

The unit of analysis is the Wikipedia content article: a structured encyclopaedic entry on a specific topic, written and continuously revised by volunteer editors [30]. What matters analytically for this thesis is not just the text but the wikilinks each article contains, connecting one article to another. These links are not incidental navigation aids. They encode editorial judgments about which concepts, persons, institutions, and events are related enough to warrant a connection. The English Wikipedia article on Clarice Phelps [31] (for an excerpt see **Figure 1A**), for example, links to nuclear chemistry, the chemical element tennessine, and the University of Texas at Austin. These connections do more than place her biographically. They locate her within an intersecting network of scientific knowledge, professional communities, and institutional affiliations. Studied at scale, such networks reveal how cultural knowledge is organized: which domains cluster together, which figures or concepts bridge fields, and how different knowledge communities relate over time.

Constructing a network from this structure requires a deliberate representational choice. As Ahnert et al. [2] caution, nothing is naturally a network: the network is an

[4] As of the time of writing (April 20, 2026) the actual number of content articles in the English Wikipedia is 7,171,064 [29].

abstraction, a framework imposed on a complex system in order to make it tractable for analysis. In the representation used here each Wikipedia article becomes a node and each wikilink becomes a directed edge (i.e., an arrow pointing from the linking article to the linked one). Directionality matters: a link from the article on Clarice Phelps to the article on the University of Texas at Austin does not imply the existence of the reciprocal link (see **Figure 1B**). This abstraction inevitably loses some information. The wikilink between Phelps and the university does not specify whether she studied or worked there, or how significant that connection was. For this thesis, that loss is acceptable. The analysis focuses not on the precise nature of individual links but on the large-scale structural patterns that emerge across the network as a whole. Those patterns are invisible at the level of any single article. They only become legible when the full network is considered.

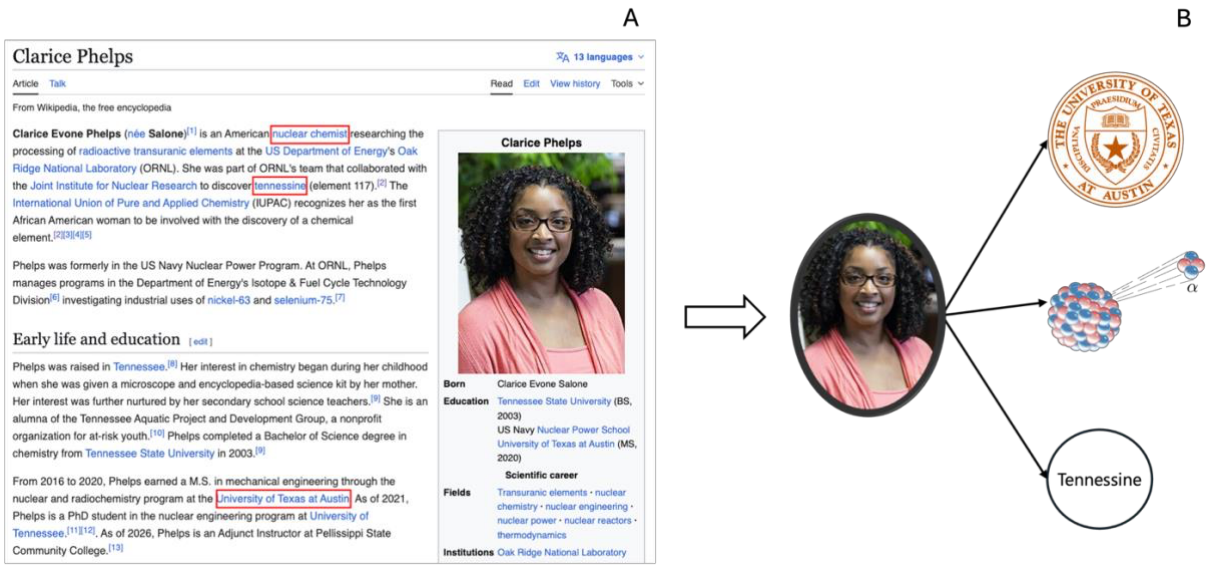


Figure 1 (A) The English Wikipedia article on Clarice Phelps, retrieved in April 2026 [31]. Red boxes highlight a selection of hyperlinks (wikilinks) embedded in the article text. **(B)** The directed network derived from the links highlighted in (A). Each Wikipedia article is represented as a node, and each directed edge indicates the presence of a wikilink from one article to another. The directionality of the edges reflects the asymmetric nature of hyperlinks: the existence of a link from the article on Clarice Phelps to the article on the University of Texas at Austin does not imply the existence of the reciprocal link.

This analytical orientation rests on the broader assumption presented in *section 1.3*: that meaning does not reside solely in individual entities but also emerges from the

patterns of relations that connect them [32, 33]. What becomes analytically important is not any single node in isolation, but the relational structure that nodes collectively form. This is the commitment that underlies the "network turn" in the humanities, which argues that network analysis provides a shared conceptual language for examining complex cultural and social systems, while also demanding critical reflection on how such representations are constructed and by whom [2, 8, 34]. The analytical value of operating at this scale is that it suspends the assumption that only already-recognised relationships are worth examining. Schich et al. [8] , for example, showed that aggregate cultural data can surface unexpected correlations between events, actors, and historical periods that specialised, small-scale studies would be unlikely to detect not because those correlations were unknown, but because they were only visible at the population level. Applying this logic to Wikipedia, the wikilink network makes it possible to identify which cultural entities function as structural hubs through which knowledge circulates, which remain peripheral or isolated, and how these positional patterns shift across historical time.

Building on this, Ahnert et al. [2] argue that such an approach is generalisable beyond any single corpus. When cultural entities are understood not merely as nodes but as carriers of ideas, practices, and forms of knowledge, network analysis becomes a tool for mapping how those elements are distributed, clustered, and connected at scale. In the context of this thesis, analysing Wikipedia's network does more than describe how articles relate to one another in a technical sense. It makes it possible to examine which cultural entities are structurally central to Wikipedia's organisation of knowledge, how the connectivity of biographical, scientific, and intellectual entities varies across historical periods, and which relationships would remain invisible if articles were examined individually rather than as components of a large, interconnected system. This supports the central claim of the thesis: that Wikipedia’s network encodes culturally meaningful patterns of connectivity that can be recovered, quantified, and interpreted through computational methods. *Cultural connectivity*, as used throughout

this thesis, refers to the structural patterns of association between cultural entities as encoded in Wikipedia's network. More precisely, it denotes which entities occupy dense, central positions within that network, which function as bridges between otherwise separate domains, and how these positional patterns shift across historical periods and knowledge communities. Moreover, the networks analysed in this thesis are understood, as mentioned earlier (see *section 1.3*), as *cultural networks* which require the relationships between people to be mediated through shared cultural entities. In the context of this thesis, that mediation is operationalized through Wikipedia's wikilink structure: two cultural entities are considered related not because one article links directly to the other, but because they share common linking contexts within the encyclopaedia's broader network of knowledge. However, before those patterns can be analysed, a prior question must be addressed: what motivates the use of Wikipedia for this type of analysis?

1.4.1. Why Wikipedia’s network and no other alternatives?

Wikipedia’s network is the most analytically appropriate resource for examining large-scale patterns of cultural connectivity, and a brief comparison with the main alternatives clarifies why. This claim rests on two criteria: the heterogeneity of the entities a resource covers, and the density and variety of the relationships it encodes. Considered against these criteria, the principal alternatives each fall short in ways that are structurally significant.

Wikidata [35], the formal knowledge base underlying many Wikipedia infoboxes, organises entities through explicitly typed relationships such as *instance of*, *part of*, and *influenced by*. This typing is analytically useful for targeted queries, but it captures only the small subset of relationships that editors have defined through controlled vocabularies. The much denser web of connections embedded in Wikipedia's article prose falls almost entirely outside Wikidata's scope or appears there only selectively and unevenly. DBpedia [36] faces a structurally comparable limitation: it extracts

structured data from Wikipedia's infoboxes and semi-structured markup, but leaves the far richer network of in-text wikilinks largely unaddressed. Because the in-text wikilink network is precisely the analytical object of this thesis, neither Wikidata nor DBpedia offers a viable substitute.

Scholarly citation networks present a different kind of limitation. They are restricted to a specific type of cultural entity, primarily published academic works and their authors, and to a single, formalised relationship type: citation. This restriction makes them poorly suited to the kind of analysis pursued here, which requires a network spanning heterogeneous entities, including people, ideas, scientific concepts, institutions, artworks, and historical events, alongside the varied associative relationships between them.

To the best of my knowledge, Wikipedia's network is the only resource that satisfies both criteria at scale. It is encyclopaedic in scope, culturally heterogeneous in its entities, and the product of sustained collective editorial judgement rather than automated extraction. Its links are not formally typed, which means some semantic precision is sacrificed. However, for the purposes of detecting large-scale structural patterns of cultural connectivity, coverage and heterogeneity matter more than formal completeness. To understand deeper the patterns I analyse, the next section will answer what Wikipedia recognises as knowledge in the first place, and through which institutional mechanisms that recognition is granted.

1.4.2. What is considered knowledge in Wikipedia?

> "The bulk of the world's knowledge is an imaginary construction".
>
> Helen Keller, The World I Live In [37]

The previous section established that Wikipedia's network encodes culturally meaningful patterns of connectivity. This raises a question: meaningful according to

whom, and on what basis? What enters the network is not determined by cultural significance alone; it is governed by a normative framework that decides what counts as knowledge worth recording. This section examines that framework, situating Wikipedia epistemologically before analysing the editorial principles through which inclusion and exclusion operate in practice.

The concept of collective memory provides the appropriate theoretical entry point, because it captures the central tension between knowledge as an objective record and knowledge as a selective, socially constructed account of what a community has chosen to preserve. In social theory, collective memory refers to shared understandings of the past that are produced and maintained by groups rather than by isolated individuals [38, 39]. Maurice Halbwachs, who coined the term, argued that remembering is always shaped by social frameworks (e.g., cultural norms, institutions, and shared narratives) that determine what groups choose to remember, how they interpret events, and which accounts are sustained over time [40]. Collective memory is therefore not a neutral archive; it is selective and oriented toward present concerns, foregrounding certain events and actors while marginalising others [41, 42]. Crucially, it is also dynamic: rather than accumulating as a stable deposit, collective memories are continuously reconstructed through communication, media, and commemorative practices, making them contested and subject to revision [43].

Jan Assmann's distinction between cultural and communicative memory refines this picture in ways directly relevant to Wikipedia. Cultural memory is stabilised in durable media such as texts, archives, and institutional practices, and is characterised by formalisation and long-term transmission [44, 45]. Communicative memory, by contrast, circulates in everyday interactions, spans living generations, and is more fluid and revisable [42, 44, 45]. Digital technologies have substantially transformed both forms, expanding the volume of material that can be preserved and the speed at which it circulates [46, 47]. However, as Haux et al. [48] observe, digital memory systems do not map cleanly onto either category: they combine large-scale storage with

structured filtering that determines not only what is preserved, but what is findable, linkable, and therefore culturally operative.

Wikipedia occupies a hybrid position within this landscape. It lacks the curatorial authority of a museum or the legal deposit function of a national archive, yet its content depends on precisely those institutions, drawing on scholarly publications and archival sources as its primary evidence base [45, 49]. At the same time, it is produced through a decentralised, ongoing process of editorial negotiation in which contributors collectively decide what should be included, how topics are framed, and how articles are interconnected [49]. Wikipedia thus combines the formalisation characteristic of cultural memory, through its reliance on citeable sources, with the dynamism and contestation characteristic of communicative memory, through the continuous renegotiation of its content. It is not a static record of what is known, but an evolving representation of what a large, distributed community has agreed is worth knowing.

This characterisation has direct methodological consequences. If Wikipedia functions as a normatively regulated, collectively constructed form of digital cultural memory, then its editorial policies are not incidental administrative rules; they are the mechanisms through which the boundary between knowledge and non-knowledge is drawn. Those policies cluster around four interlocking principles: verifiability, notability, reliance on secondary sources, and the requirement of a neutral point of view.

Verifiability requires that all content be supported by reliable, published sources that readers can independently consult [50]. Material that lacks such support cannot be included, regardless of its factual accuracy or cultural significance [51, 52]. Original research (i.e., the proposal of new interpretations or theories) is explicitly prohibited, anchoring Wikipedia's content to what has already been recognised and published elsewhere [51]. *Notability* determines whether a topic warrants a dedicated article at

all. A topic is considered notable if it has received significant coverage in reliable sources that are independent of the subject [50]. This principle functions as the primary mechanism of inclusion and exclusion: it is the editorial filter that decides whether a cultural entity gains a node in the network or remains absent from it. *Reliability*, in this context, refers to sources with established editorial standards, such as peer-reviewed publications and reputable news organisations [53]; independence requires that those sources have no direct interest in the subject they cover [54]. *Secondary sources and neutrality* reinforce this structure. Wikipedia prioritises secondary sources which means that the attention falls on sources that provide analysis and interpretation of events rather than first-hand accounts [52], and requires that all content represent significant viewpoints fairly, as they appear in the published record [53, 55]. Together, these principles mean that what enters Wikipedia is not simply what is true, but what has been recognised, published, and independently documented by sources that meet institutional standards.

Taken together, these four principles constitute a filter that operates prior to any analysis conducted on the network: they determine which cultural entities are represented, how prominently, and in relation to which others. This has a direct implication for this thesis. The patterns of connectivity recovered from Wikipedia's network do not reflect an unmediated map of cultural knowledge; they reflect a map of knowledge as it has been certified, formalised, and selectively preserved by a specific epistemic regime; this regime operates at the intersection of the editorial community's collective judgements about what is worth including and the formal constraints imposed by Wikipedia's editorial guidelines. Understanding this regime is therefore not a preliminary formality but an interpretive condition because it shapes what the network can and cannot reveal, and what must be kept in view when patterns of inclusion and connectivity are analysed. The following section examines how these knowledge criteria translate directly into network structure: which entities gain nodes,

which connections form edges, and which cultural entities remain structurally absent from the system altogether.

1.4.3. How Inclusion Shapes the Network of Wikipedia

The four principles examined earlier (i.e., verifiability, notability, secondary sourcing, and neutrality) as well as the biases that each editor carries (see *Limitations*) do not only determine the textual content of individual articles. They also determine the topology of the wikilink network itself. Because a content article can exist only if its subject meets notability and verifiability standards, any topic that fails these criteria is absent from the network not merely as a gap in coverage but as a structural condition: it has no node, and therefore no edges can form to or from it. The network's connectivity patterns are thus shaped upstream, before any links are created, by a filtering process that is normative rather than neutral. What the network makes visible is precisely what this filtering regime has admitted; what it cannot reveal is what the regime has excluded.

The case of Donna Strickland illustrates this mechanism concretely. Prior to October 2018, the English Wikipedia contained no article on her, because editors determined that the available independent coverage of her work did not yet satisfy notability standards [56]. In network terms, she was structurally absent: no node existed, and no wikilink could point to or from her. This absence was not a reflection of her scientific significance because she had already made the contributions that would earn her a Nobel Prize, but of the documentation threshold that Wikipedia's editorial framework requires. When she was awarded the Nobel Prize in Physics in 2018 alongside Gérard Mourou, the resulting surge in independent coverage triggered the creation of her article, and she entered the network with links to related scientific fields, institutions, and co-recipients.[5] The analytical implication is significant: the network's apparent

[5] As of March 2026, the article of Donna Strickland in the English Wikipedia has an out-degree of 66 (wikilinks to other Wikipedia content articles) and an in-degree of 56 (the number of wikilinks that her article receives from other Wikipedia content articles).

completeness at any given moment conceals a prior layer of selectivity. Structural centrality in the network reflects not only cultural significance in the world but the degree to which that significance has been recognised, documented, and certified by sources meeting Wikipedia's institutional standards. However, it should be noted that her case points to a further layer of complexity. Because Wikipedia's notability criteria depend on coverage in secondary sources, the network also inherits the selectivity of the broader scholarly and journalistic record; a record in which women have historically been underrepresented relative to their contributions and which record affects what enters Wikipedia [57, 58].[6]

Two further qualifications follow from this analysis and bear directly on the empirical approach adopted in this thesis. First, it is worth being clear about what this analysis observes and what it does not. The thesis does not look directly at how Wikipedia's knowledge is made (the back-and-forth of editorial debates, contested revisions, and talk-page arguments about whether someone deserves an article); rather, it examines the outcome of those processes: the network of articles and links as it existed at a fixed point in time. Following this line of thought, the structure under study encodes the accumulated effect of thousands of editorial decisions, but it does not show those decisions happening. The cultural memory framework introduced above is used here because it helps make sense of what kind of information the resulting structure is; a sedimented record of what the community of Wikipedia editors judged worth preserving, connecting, and certifying.

Second, the empirical focus is on the English Wikipedia specifically. This choice is motivated by its scale — as the largest language edition, with approximately 7.2 million articles [29] and by the particular character of its contributor base: because English functions as a global lingua franca of the internet, the English Wikipedia attracts editors from a wide range of geographical and cultural backgrounds [45], producing a breadth

[6] It is worth mentioning that a recent study showed that after 2022 the representation gap has been reversed in biographical pages of women biologists affiliated with United States universities [59].

of coverage that makes it the most suitable version for a large-scale analysis of cultural connectivity across historical periods. Neither of these conditions eliminates the selectivity documented above; they constrain it in specific ways that the analysis keeps in view.

These considerations establish the interpretive frame within which the patterns identified in the following chapters should be read. The wikilink network analysed in this thesis is not a neutral map of cultural history; it is a structured, normatively regulated representation of knowledge as a specific epistemic community — Wikipedia's collective of editors — has chosen to record and connect it. Its patterns of connectivity reflect the cumulative effect of editorial judgements about what is notable, verifiable, and relevant, mediated by the institutional standards of secondary scholarship and shaped by the known biases of the contributing community (see *Limitations*). The analytical results should therefore be interpreted not as direct evidence of historical reality, but as evidence of how that reality has been selectively encoded in a large, collaboratively constructed digital knowledge system. This distinction is not a limitation to be apologised for; it is the interpretive condition that makes the analysis possible and meaningful. With these conceptual and methodological foundations in place, the thesis addresses three research questions, each organised as a chapter linked to its corresponding paper and each addressing a distinct but interrelated dimension of cultural connectivity in Wikipedia's network.

1.5. Why Studying Wikipedia as a Cultural Network Matters

Understanding the significance of studying Wikipedia as a cultural network requires zooming out from its everyday function as an encyclopaedia. What is important for this thesis is the relational structure that connects Wikipedia articles to one another (i.e., the wikilinks). This relational structure is one of the largest collaboratively constructed records of how a distributed community of contemporary editors organises, connects, and represents cultural knowledge. When this network is studied

as a cultural network in which relationships between people are mediated by cultural entities such as ideas, institutions, and artefacts [28], it reveals the significance of this study.

To begin with, Wikipedia's network provides a basis for examining how cultural knowledge is structured at a scale that no single disciplinary tradition is designed to address on its own. For example, a historian of science may trace influence networks among 17th- century natural philosophers. A literary scholar may map intertextual references across a national canon. Even though studies of this type are valuable, they will illuminate patterns within their own domains. However, they will face an obstacle to capture connections that run between domains. Wikipedia's network does not face this constraint: it links articles on scientific discoveries, political movements, literary works, individuals, and artworks within the same connective structure. By analysing the connectivity patterns that emerge across millions of articles spanning centuries and fields, it becomes possible to ask a different kind of question. Among the questions this enables are the following: which cultural entities function as bridges between otherwise separate fields? Which domains remain isolated from one another? These questions, as stated above, are difficult to address by examining any single domain in isolation. They require the kind of cross-domain, large-scale perspective that an analysis of the cultural network of Wikipedia offers and affords.

Second, as shown earlier, Wikipedia's content is governed by a set of normative editorial principles and the biases that each editor carries. Because of these factors, the network does not passively mirror cultural reality. Instead, it reflects a particular and normatively regulated representation of that reality. For this reason, studying Wikipedia's cultural network as a collaboratively constructed knowledge system also offers insight into the patterns of inclusion and exclusion that the editorial community has embedded in the network.

1.6. Research Questions

The preceding sections have established the conceptual foundations. Wikipedia is introduced in this thesis as a collaboratively constructed form of cultural network, whose internal link network provides a structural, large-scale record of how cultural entities (i.e., people, ideas, cultural institutions, etc.) are associated with one another and what the editorial community decided to include or exclude. The analytical task at hand is to use this network as a data source from which connectivity patterns between cultural entities can be identified, quantified, and interpreted through humanistic interpretive frameworks. Three research questions which are presented as chapters linked to corresponding papers, organise this task. Each addresses a distinct but interrelated aspect of the overall project, and together they trace a progression from methodological proof of concept, through the development of the necessary computational infrastructure, to a large-scale empirical analysis.

The research questions (RQ) this thesis answers are the following:

- **RQ1**: Can a network-based analysis of Wikipedia's wikilink structure recover meaningful connectivity patterns between cultural entities in a given historical period?
- **RQ2**: How can a scalable, multilingual pipeline for extracting text and wikilink networks from Wikipedia's content articles be designed and implemented?
- **RQ3**: What temporal and semantic patterns emerge in the biographical subnetwork of the English Wikipedia when analysed across twenty-five centuries?

Bibliography

1. Miccio LA, Agapitos P, Gamez-Perez C, et al (2025) Wikipedia as a cultural lens: a quantitative approach for exploring cultural networks. Humanit Soc Sci Commun 12:462. https://doi.org/10.1057/s41599-025-04772-5

2. Ahnert R, Ahnert SE, Coleman CN, Weingart SB (2020) The Network Turn: Changing Perspectives in the Humanities, 1st ed. Cambridge University Press

3. Edelmann A, Wolff T, Montagne D, Bail CA (2020) Computational Social Science and Sociology. Annu Rev Sociol 46:61–81. https://doi.org/10.1146/annurev-soc-121919-054621

4. Manovich L (2020) Cultural analytics. The MIT Press, Cambridge, Massachusetts; London, England

5. Lazer D, Pentland A, Adamic L, et al (2009) Computational Social Science. Science 323:721–723. https://doi.org/10.1126/science.1167742

6. Salah AA, Manovich L, Salah AA, Chow J (2013) Combining Cultural Analytics and Networks Analysis: Studying a Social Network Site with User-Generated Content. Journal of Broadcasting & Electronic Media 57:409–426. https://doi.org/10.1080/08838151.2013.816710

7. Romein CA, Kemman M, Birkholz JM, et al (2020) State of the Field: Digital History. History 105:291–312. https://doi.org/10.1111/1468-229X.12969

8. Schich M, Song C, Ahn Y-Y, et al (2014) A network framework of cultural history. Science 345:558–562. https://doi.org/10.1126/science.1240064

9. So RJ (2021) Introduction: Contemporary Culture After Data Science. jca 6:1268. https://doi.org/10.22148/001c.22335

10. Cain WE (2018) British and American New Criticism. In: Richter DH (ed) A Companion to Literary Theory, 1st ed. Wiley, pp 9–23

11. Bode K (2017) The Equivalence of "Close" and "Distant" Reading; or, Toward a New Object for Data-Rich Literary History. Modern Language Quarterly 78:77–106. https://doi.org/10.1215/00267929-3699787

12. Gallop J (2007) The Historicization of Literary Studies and the Fate of Close Reading. Profession 181–186

13. Su T (2009) On Close Reading from the Perspective of Rhetoric. ASS 5:p150. https://doi.org/10.5539/ass.v5n8p150

14. Ohrvik A (2024) What is close reading? An exploration of a methodology. Rethinking History 28:238–260. https://doi.org/10.1080/13642529.2024.2345001

15. Middleton P (2005) Distant reading: performance, readership, and consumption in contemporary poetry. the University of Alabama press, Tuscaloosa

16. Moretti F (2000) Conjectures on World Literature. NLR 2:54–68. https://doi.org/10.64590/hxj

17. Moretti F (2005) Graphs, maps, trees: abstract models for a literary history, 1. publ. Verso, London New York

18. JOCKERS ML (2013) Macroanalysis. University of Illinois Press

19. Underwood T (2017) A Genealogy of Distant Reading. Digital Humanities Quarterly 11:. https://doi.org/10.63744/ktwb2atwsbep

20. Sula CA, Hill HV (2019) The early history of digital humanities: An analysis of Computers and the Humanities (1966–2004) and Literary and Linguistic Computing (1986–2004). Digital Scholarship in the Humanities fqz072. https://doi.org/10.1093/llc/fqz072

21. Miccio LA, Gámez-Pérez C, Suárez JL, Schwartz GA (2022) Mapping the Networked Context of Copernicus, Michelangelo, and Della Mirandola in Wikipedia. Advs Complex Syst 25:2240010. https://doi.org/10.1142/S0219525922400100

22. Mr. Malay Das (2025) Cultural Studies and Digital Humanities. https://doi.org/10.5281/ZENODO.15657666

23. Wallace M, Poulopoulos V, Antoniou A, López-Nores M (2023) An Overview of Big Data Analytics for Cultural Heritage. BDCC 7:14. https://doi.org/10.3390/bdcc7010014

24. Rahman AM (2025) Digital Humanities: Computational Methods For Research. International Research Journal of Arts and Social Sciences 13:114. https://doi.org/10.14303/2276-6502.2025.114

25. Dondero MG (2019) Visual semiotics and automatic analysis of images from the Cultural Analytics Lab: How can quantitative and qualitative analysis be combined? Semiotica 2019:121–142. https://doi.org/10.1515/sem-2018-0104

26. Drucker J (2020) Visualization and Interpretation: Humanistic Approaches to Display. The MIT Press

27. Ahnert R, Ahnert SE, Coleman CN, Weingart SB (2020) The Network Turn: Changing Perspectives in the Humanities, 1st ed. Cambridge University Press

28. Suarez J-L, McArthur B, Soto-Corominas A (2015) Cultural Networks and the Future of Cultural Analytics. In: 2015 International Conference on Culture and Computing (Culture Computing). IEEE, Kyoto, Japan, pp 95–98

29. Statistics - Wikipedia. https://en.wikipedia.org/wiki/Special:Statistics. Accessed 20 Apr 2026

30. Kopf S (2022) 2.1.1. The Article Page. In: A Discursive Perspective on Wikipedia: More Than an Encyclopedia? Springer Nature, Switzerland, pp 30–34

31. (2026) Clarice Phelps. Wikipedia

32. Gabella M (2019) Cultural Structures of Knowledge from Wikipedia Networks of First Links. IEEE Trans Netw Sci Eng 6:249–252. https://doi.org/10.1109/TNSE.2018.2812788

33. Easley D, Kleinberg J (2010) Networks, crowds, and markets: reasoning about a highly connected world. Cambridge University Press, Cambridge

34. Goldfarb D, Merkl D, Schich M (2015) Quantifying Cultural Histories via Person Networks in Wikipedia

35. Wikidata. https://www.wikidata.org/?uselang=en. Accessed 21 Apr 2026

36. (2025) Home. In: DBpedia Association. https://www.dbpedia.org/. Accessed 21 Apr 2026

37. Keller H (2004) The World I Live In. NYRB Classics

38. Roediger HL, Abel M (2015) Collective memory: a new arena of cognitive study. Trends in Cognitive Sciences 19:359–361. https://doi.org/10.1016/j.tics.2015.04.003

39. Graus D, Odijk D, De Rijke M (2018) The birth of collective memories: Analyzing emerging entities in text streams. Asso for Info Science & Tech 69:773–786. https://doi.org/10.1002/asi.24004

40. Halbwachs M (1992) On collective memory, Nachdr. University of Chicago Press, Chicago

41. Nora P (1989) Between Memory and History: Les Lieux de Mémoire. In: Representations. pp 7–24

42. Yasseri T, Gildersleve P, David L (2022) Collective memory in the digital age. In: Progress in Brain Research. Elsevier, pp 203–226

43. Ferron M, Massa P (2014) Beyond the encyclopedia: Collective memories in Wikipedia. Memory Studies 7:22–45. https://doi.org/10.1177/1750698013490590

44. Assmann J (2011) Communicative and Cultural Memory. In: Meusburger P, Heffernan M, Wunder E (eds) Cultural Memories. Springer Netherlands, Dordrecht, pp 15–27

45. Pentzold C (2009) Fixing the floating gap: The online encyclopaedia Wikipedia as a global memory place. Memory Studies 2:255–272. https://doi.org/10.1177/1750698008102055

46. Mößner N, Kitcher P (2017) Knowledge, Democracy, and the Internet. Minerva 55:1–24. https://doi.org/10.1007/s11024-016-9310-0

47. Gunn H, Lynch MP (2021) The Internet and Epistemic Agency. In: Applied Epistemology. Oxford University Press, Oxford, United Kingdom; New York, NY, pp 390–409

48. Haux DH, Maget Dominicé A, Raspotnig JA (2021) A Cultural Memory of the Digital Age? Int J Semiot Law 34:769–782. https://doi.org/10.1007/s11196-020-09778-7

49. Kopf S (2022) A Discursive Perspective on Wikipedia: More than an Encyclopaedia? Springer International Publishing, Cham

50. (2026) Wikipedia:Notability. Wikipedia

51. (2026) Wikipedia:Verifiability. Wikipedia

52. (2026) Wikipedia:No original research. Wikipedia

53. (2026) Wikipedia:Reliable sources. Wikipedia

54. (2026) Wikipedia:Independent sources. Wikipedia

55. (2026) Wikipedia:Neutral point of view. Wikipedia

56. Erhart E (2018) Why didn't Wikipedia have an article on Donna Strickland, winner of a Nobel Prize? In: Diff. https://diff.wikimedia.org/2018/10/04/donna-strickland-wikipedia/. Accessed 16 Apr 2026

57. Tripodi F (2023) Ms. Categorized: Gender, notability, and inequality on Wikipedia. New Media & Society 25:1687–1707. https://doi.org/10.1177/14614448211023772

58. Lemieux ME, Zhang R, Tripodi F (2023) "Too Soon" to count? How gender and race cloud notability considerations on Wikipedia. Big Data & Society 10:20539517231165490. https://doi.org/10.1177/20539517231165490

59. Alvarez-Ponce D, Iyengar N (2026) Monitoring the gender gap in the coverage of biology professors on Wikipedia. Proceedings of the Royal Society B: Biological Sciences 293:20252566. https://doi.org/10.1098/rspb.2025.2566

2. Methodology

As mentioned in the *Introduction*, the goal is to use English Wikipedia's wikilinks network as a quantitative cultural lens through which patterns of connectivity between cultural entities[7] can be systematically examined. By treating these links as measurable connections, it becomes possible to systematically examine patterns of cultural connectivity across time, and field. Therefore, the studies presented in *section 3.1* and *section 3.2* share Wikipedia as their data source. Each one of them operates at a different scale and follow different methodological techniques: network analysis and distributional semantics. Together, these two frameworks provide both the structural and the semantic tools needed to study Wikipedia's cultural network of information. More specifically, focusing on the individual studies themselves, *section 3.1* works at the micro-to-meso scale, concentrating on a single historical period. It reconstructs the interdisciplinary cultural network that connected Art, Science, and Philosophy during the 17th century CE. To do so, it employs a structural relatedness metric inspired by the Normalized Google Distance, a method that infers closeness from linking patterns, and uses bootstrapping, a resampling technique that estimates the stability of results by repeatedly drawing subsets from the data. *Section 3.2*, by contrast, opens the scale considerably. It spans twenty-five centuries of recorded history and draws on nearly two million Wikipedia biographical articles about historically notable individuals. Rather than focusing on a single period, it combines network topology (the mathematical study of how nodes and edges are arranged and connected) with semantic embeddings (numerical representations that capture meaning through position in a n-dimensional space) in order to uncover long-range temporal patterns that would be invisible at smaller scales. The following sections present the methodology of each study in turn, beginning with the most granular and progressing toward the most expansive.

[7] For a definition of cultural entities in the context of this thesis refer to *Introduction*.

2.1. Introduction

Network analysis has become one of the primary tools through which the field of digital humanities and cultural analytics approach culture at scale which, as stated in the *Introduction*, has been crystallised through the "network turn" [1–5]: the idea that interconnected information (e.g., who or what is connected to whom, and when) carries meaningful cultural information.

As introduced earlier, the networks employed in this thesis belong to a specific category within this broader tradition. They are cultural networks: graphs in which nodes represent cultural entities such as people, ideas, artworks, and events, and in which edges encode meaningful associations between them. However, these networks require careful interpretation. Because Wikipedia is written and maintained by contemporary editors, each connection reflects two overlapping layers of information rather than a transparent record of the past. The first is a historical layer: documented facts about past people and events. The second is an editorial layer: the choices that present-day editors have made about what to record, emphasise, and connect. It is this second layer that makes Wikipedia an artefact of cultural memory — a selective, constructed record of what a particular community of editors, writing at a specific moment in time, judged to be worth preserving and relating.

Both layers introduce their own analytical challenges. The historical layer is uneven: information about the distant past is sparser, less consistent, and more prone to gaps than information about recent centuries. The editorial layer introduces representational issues, reflecting the demographic and cultural composition of Wikipedia's contributor base rather than any neutral standard of historical significance. These are not incidental problems but structural features of the data source that affect the methodological choices made in this thesis. As detailed in the sections that follow, the analytical framework adopted here is designed to account for both layers of bias, with the aim of producing a more representative and interpretively robust analysis.

2.2. Period characterization[8]

2.2.1. Structural Relationships using the Normalised Google Distance

Network analysis has emerged as one of the primary methodological tools through which digital humanities and cultural analytics examine culture at scale. As noted in the *Introduction*, this methodological shift has been theorised as the "network turn" [1–5]. Most of the research works that use Wikipedia's internal links network are using the direct connections between articles. For instance Aragon et al. analyse biographical social networks by mapping links between biographies across 15 language editions to identify central historical figures and cross-cultural similarities [7]. Another study focuses on the "First Link Network," following the first wikilink in the body of 4.7 million articles to reveal a flow from specific topics toward general, foundational concepts like "Philosophy" [8]. This structural analysis extends to identifying cultural hierarchies, demonstrating how different linguistic editions reflect distinct intellectual heritages, such as the Western emphasis on "Science" versus the East Asian focus on concepts like "Human" and "Earth" [9]. Moreover, within the field of digital history, researchers use the wikilink structure between articles on historical figures to quantify their importance and explore the temporal relationships between link architecture and article popularity [10]. Additionally, the growth of scientific knowledge is examined by treating linked articles as concept networks, showing that science in Wikipedia progresses by identifying and filling "knowledge gaps" or topological cavities over time [11]. Lastly, a recent multicultural analysis of philosophers employs the directed network of hypertext links across nine languages to rank intellectual influence, using

[8] The methodology presented in *Section 2.2* has been published as part of a peer-reviewed article with the title "Wikipedia as a cultural lens: a quantitative approach for exploring cultural networks" [6].

algorithms like PageRank, and to uncover hidden connections between schools of thought [12].

The analyses presented above are mainly based on explicit, or direct, connections between articles, but this methodological choice systematically obscures implicit cultural relationships. Hence, this limitation motivates a shift in how relationships between Wikipedia articles are conceptualized. Instead of regarding the network as a mere set of direct links, the present study introduces the notion of the *structural relationship*, which captures associations between articles even if they are not directly linked by a wikilink. The underlying idea is that if two Wikipedia articles frequently appear close to each other in the network, sharing most of the articles that link them and most of the articles they link to, then the network itself indicates that these two topics are conceptually the same. This signal persists regardless of whether a direct link between the two articles exists (see **Figure 2**). In other words, two topics are considered conceptually close if they consistently appear within the same relational context. This conceptual proximity is assumed to reflect the cumulative editorial judgment of thousands of independent contributors who, without necessarily intending to connect two topics directly, have repeatedly situated them within the same relational context. Therefore, instead of viewing the Wikipedia network as merely a collection of direct connections, we incorporate the idea of structural relationship between nodes.

It should be noted that this idea of implicit relationships between articles is not new. Studies by [13, 14] have examined Wikipedia's cultural network through the lens of implicit relationships, demonstrating that treating indirect connections reveals meaningful patterns absent from direct-link analysis. Nevertheless, those studies applied their method to a very small set of specific nodes, rather than to a full characterisation of a cultural domain. As opposed to this, the analysis carried out later, as shown, brings together multiple cultural entities by averaging their connectivity patterns.

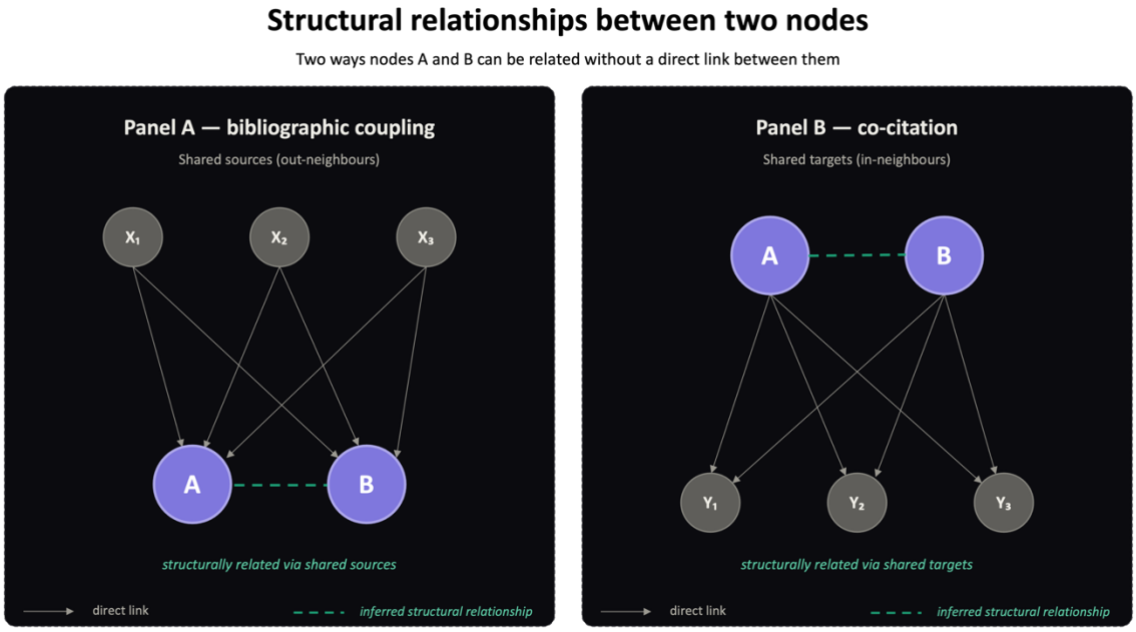


Figure 2 Structural relationship between elements in a complex network. The example illustrates how the Normalised Google Distance (NGD) detects structurally similar nodes even if there is not direct links between them.

To detect and quantify this type of structural relationship, we use a metric inspired by the Normalised Google Distance (hereafter NGD) [15, 16]. The measure captures the overlap in link neighbourhoods. Formally, the NGD between two articles a and b is defined as [15]:

$$d_{in/out}(\mathrm{a},\mathrm{b}) = \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))}$$

Where a and b denote the two articles of interest; A and B represent the set of Wikipedia articles that link to, or are linked from, a and b respectively, depending on whether the in-degree (d_{in}) or the out-degree (d_{out}) is being computed; and W denotes the total number of nodes (Wikipedia articles) in the network. When $|A \cap B| = 0$, the distance is undefined and treated as infinite, reflecting the complete absence of any shared neighbourhood between the two articles.

The interpretation of the formula is the following: The numerator compares the shared neighbourhood to the larger of the two individual neighbourhoods, expressed as a log-ratio that grows as the overlap shrinks. The denominator then rescales this value

against the size of the full network, so that the resulting distance reflects relative rather than absolute overlap. Consequently, two articles whose neighbourhoods overlap substantially receive a small distance, whereas two articles whose neighbours barely intersect receive a large one. A further property follows from this construction: the measure responds not only to the raw number of shared neighbours, but also to the size of each neighbourhood. Ten shared neighbours, for instance, carry a very different meaning for two sparsely connected articles than for two densely connected ones.

For each pair of articles a and b, two distances are computed separately. The first, $d_{in}(a, b)$, is based on the set of articles that link to both a and b, and therefore captures how similarly the rest of the network refers to them. The second, $d_{out}(a, b)$, is based on the set of articles that both a and b link to and therefore captures how similarly their authors situate them within the broader conceptual landscape. The final distance $d(a, b)$ is then defined as the harmonic mean of these two quantities. Because the harmonic mean is dominated by its smaller input, the combined distance is pulled toward whichever direction shows the stronger neighbour overlap. Consequently, a pair of articles which share many neighbours align closely in even one direction receives a relatively low combined distance, which makes the measure responsive to partial structural relatedness. Finally, since the theoretical distance $d(a, b)$ can range from 0 to infinity, which makes raw values hard to compare and interpret. To address this, the distance is mapped onto a bounded similarity score in the range $[0,1]$ using the transformation: $r(a, b) = \exp(-d(a, b))$. This way, identical items (distance 0) yield a similarity of 1, while increasingly dissimilar items (large distances) approach 0.

This metric offers several advantages over relying on direct links alone, both of which matter for the reliability and scope of the corresponding analysis of *section 3.1*. First, including incoming links reduces a bias that appears when only article text or outgoing links are used. Any single Wikipedia article reflects the choices of the editors who wrote it, including which topics they decided to reference. Incoming links work differently. They aggregate judgments from the broader editorial community, reflecting how

many independent contributors, writing across different articles, have chosen to place a given topic within the network. That aggregation dilutes the idiosyncratic perspective any one article carries. Second, using both incoming and outgoing links together gives a more complete picture of each article's position in the network. Outgoing links show how an article reaches toward related concepts. Incoming links show how the rest of the network locates it. Together, they reflect the structure of the encyclopaedia, not the editorial perspective of any single article.

2.2.2. People, triads and clustering

The goal is to investigate cross-field intellectual exchange in a given century. To do so, we first need to extract the biographical data from Wikipedia which will serve the basis for this analysis. These biographies will then be used to construct triads of individuals from three broad areas of knowledge creation: Art, Philosophy, and Science of the 17th century. This two-step process, extraction followed by the construction of triads, provides the structured input from which the networks analysed in the corresponding chapter (see *Results - section 3.1*) are built. The methodological choices governing each step are described in turn below.

Detection of biographies. The first step was to obtain a stable, reproducible corpus of Wikipedia articles. This was achieved by parsing a publicly available dump of the English Wikipedia, specifically a snapshot dated 29 January 2023. Wikipedia dumps are periodic static exports of the entire encyclopaedia and constitute a standard resource for large-scale computational studies of knowledge. Working from a fixed snapshot guarantees reproducibility: because the dataset is frozen at a single point in time, it does not shift as editors update articles, meaning any researcher can recover the exact same sample from the same dump. From this snapshot, individuals were identified by detecting the tag "person" in the infobox field of each article, where available.[9]

[9] The infobox is a structured metadata block appearing at the top of many Wikipedia articles, providing key factual information in a machine-readable format.

Applying this filter produced an initial sample of 501,046 persons, which forms the starting population for all subsequent classification steps.

Metadata for biographies. With the initial sample of biographical pages established, the next step was to extract the specific attributes required to identify the field to which each individual belongs and the century during which they lived. For each person, the first sentence of the corresponding Wikipedia article was parsed using regular expressions to retrieve three pieces of information: birthdate, number of internal links, and professional occupation. The birthdate enables temporal filtering, restricting the sample to the 17th century. The number of internal links serves as a proxy for an individual's relative prominence within Wikipedia's knowledge network. The profession provides the basis for field classification into Art, Philosophy, or Science. This last variable was also extracted from the first sentence of each biography rather than from the full article text, because the first sentence of a Wikipedia biography typically follows a standardised formula that explicitly names the subject's primary occupation and nationality. Relying on this sentence alone therefore yields a concise and consistent signal across hundreds of thousands of entries, without introducing the noise that full-text extraction would generate. Together, these three attributes form the biographical metadata necessary for the classification pipeline described below.

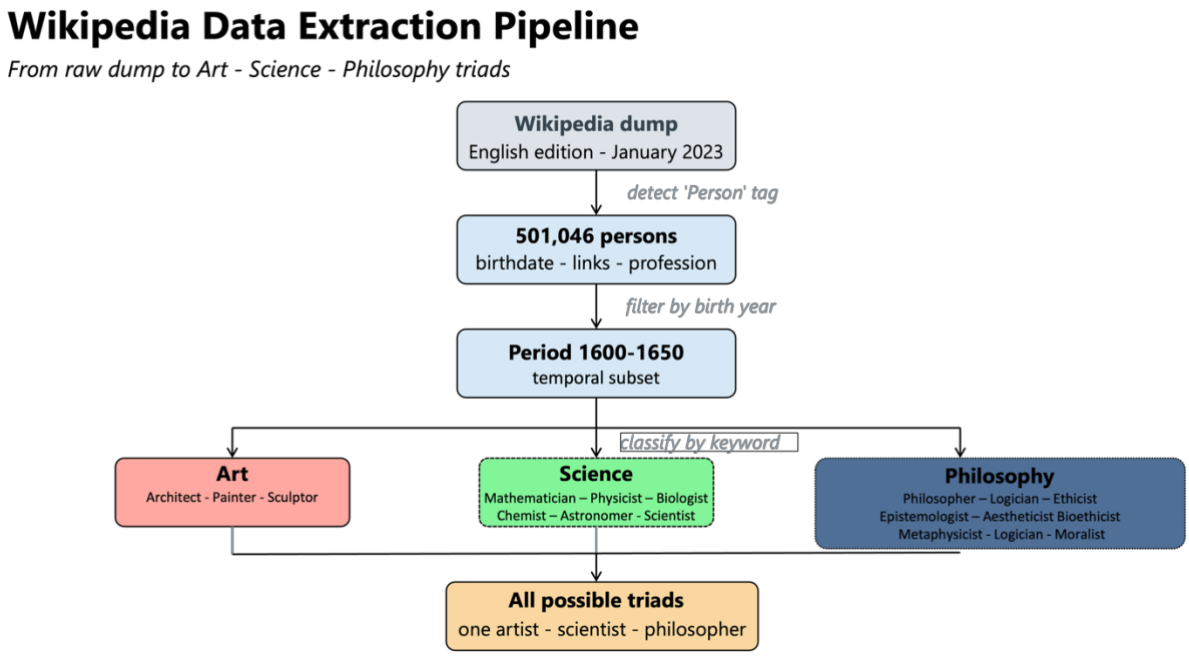


Figure 3 Schematic representation of the pipeline used to extract artists, scientists, and philosophers from Wikipedia for the period 1600-1650.

Figure 3 illustrates the full pipeline used to generate the triads. Starting from the complete set of Wikipedia pages tagged as "persons" extracted from the dump, the pipeline proceeds through a sequence of filtering and classification steps. First, the sample was narrowed to individuals born between 1600 and 1650, the historical window of interest for this study. This period was chosen because it encompasses a pivotal era in Western intellectual history, bridging the late Renaissance and the central decades of the Scientific Revolution and laying important foundations for the Enlightenment, a time when the boundaries between art, philosophy, and emerging forms of science were actively being negotiated and cross-field intellectual exchange was particularly intense [17].

Art-Philosophy-Science classification. With the sample restricted to the relevant historical period, the next step was to identify the field in which each individual belongs to given the following three fields: Art, Philosophy, or Science. This classification was performed by checking whether at least one term from a set of pre-defined keyword lists appeared among the professions extracted from each biography (see **Figure 3**). At this point, I should mention that any classification scheme of this kind involves a

degree of arbitrariness in the choice of keywords, but the approach carries three practical advantages that justify its use here. First, it is systematic and fully transparent: every classification decision follows the same rule and can be inspected and, if needed, revised. Second, it is flexible: the keyword lists can be broadened or narrowed depending on the research question, making the method readily adaptable to other historical periods or disciplinary boundaries. Third, it operates at scale without requiring manual annotation, which would be infeasible given the size of the dataset. The principal trade-off is that some individuals will inevitably be misclassified, either because their biography uses unusual occupational terminology or because they genuinely worked across multiple fields. However, given the scale of the dataset, such edge cases are unlikely to distort the aggregate patterns of interest, and the transparency of the keyword lists means that any misclassification can be identified and corrected if needed.

Moreover, we set a minimum outdegree[10] threshold of 100 wikilinks, meaning that only individuals whose Wikipedia articles contain at least 100 outgoing wikilinks were retained as nodes of the network. This threshold is set on the basis of three considerations. First, the outdegree of a Wikipedia article can serve as a proxy for an individual's relative prominence [18]. The intuition behind this idea is that more richly documented figures tend to have more outgoing links, and prominence within Wikipedia's knowledge network is a reasonable, if imperfect, indicator of historical significance. Second, articles with higher outdegree tend to have been edited more frequently and by more contributors, which is associated with higher overall article quality [19]. Third, because the size of the network generated for each person scales with that person's outdegree, a minimum threshold is necessary to ensure that the resulting networks are large enough to support statistically meaningful analysis. The value of 100 is established empirically through experience with networks drawn from

[10] The outdegree refers to the number of links that a node sends.

multiple historical period [20, 21], and represents a pragmatic balance between inclusivity and analytical reliability.

Triads and networks. Once all this information is available, the next step is to generate the triads. To do that, as mentioned earlier, I use biographical pages of people born between 1600 and 1650 that are classified as artists, scientists, or philosophers based on keywords in their biographies (see **Figure 4**). Then all possible combinations of these individuals are generated to form triads. Each triad is specifically composed of one individual from each of three pre-defined disciplines: Art, Science, and Philosophy. Using Wikipedia's internal links, I identify articles related to these individuals, forming a larger network of cultural entities (for the definition refer to *Introduction*) that includes people, objects, and ideas. An example of a triad consists of the artist Johannes Vermeer, the scientist Christian Huygens, and the philosopher Blaise Pascal. This triad-based approach permits the quantification of how knowledge flowed between different disciplines during the 17th century CE and can reveal patterns of either strong or weak connections between disciplines.

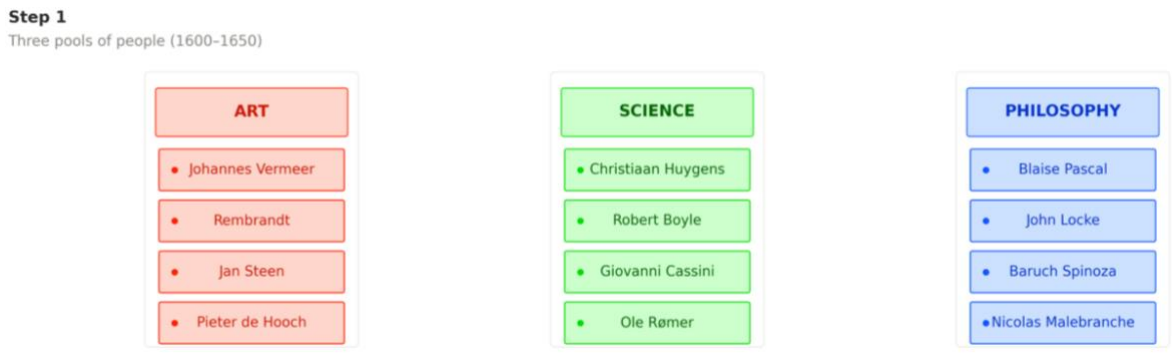


Figure 4 People's Wikipedia articles classified in one of the disciplines: Art (light orange), Science (light blue), and Philosophy (light purple).

Once the triads are established, the next step is to build a network for each one.[11] In the present context, in the network each node corresponds to a Wikipedia article, and each edge represents a wikilink connecting two articles. The goal of constructing these

[11] A network, in the technical sense used here, is a mathematical structure composed of nodes (individual entities) and edges (connections between the nodes) [22].

networks is to produce a structured, quantitative representation of the conceptual landscape surrounding each triad, capturing how the ideas, people, and topics associated with each field relate to one another.

For each person in each triad the corresponding English Wikipedia article is retrieved, along with all articles that are reachable within two wikilinks steps from it. In other words, this means collecting not only the articles directly linked from the person's page, but also the articles linked from those pages, and so on up to this second level of distance. This step was implemented to enrich but also bound the sample of the conceptual terrain surrounding each individual. For instance, for a triad such as that of Johannes Vermeer, Christiaan Huygens, and Blaise Pascal, this procedure yields a graph of 38,487 nodes and over 1.4 million edges, which illustrates the sheer scale of information that Wikipedia encodes even for a narrow selection of three historical figures.

Among the extracted links, there are a lot that do not carry substantive encyclopaedic content and they exist in Wikipedia as content pages that serve organisational purposes: redirect pages (page that automatically forward visitors to another article), disambiguation page (which list multiple possible meanings of a term), category pages, and list pages. These are removed from the graph, as they do not represent genuine conceptual entities and would introduce noise into the subsequent analysis. Redirect links are resolved prior to removal, meaning that any link pointing to a redirect is updated to point directly to the article's actual destination (e.g., "Physics" sends a link to "Einstein" (redirect) will be resolved to link to "Albert Einstein" (content article).

With the cleaned graph ready, the next step is to measure the relatedness between each seed and the other nodes in its neighbourhood. This is done by using the NGD converted to a similarity metric (see *section 2.2.1*). The directed graph of wikilinks is then converted into an undirected one, meaning that the distinction between who links to whom is discarded in favour of a simpler representation in which connections

are treated as bidirectional. Directionality on Wikipedia is not arbitrary: wikilinks tend to flow from more specific articles toward more canonical ones [8, 9, 23, 24]. For this reason, this structural asymmetry itself can be used to study reference patterns or cultural prominence through measures such as PageRank or in/out-degree. Discarding it therefore rules out a particular family of analyses. In the present case, however, this is not a problem, for two reasons. First, the relatedness measure adopted here is symmetric by construction, since NGD is derived from co-occurrence statistics that do not encode direction; keeping the graph directed after assigning symmetric weights would be inconsistent. Second, the aim of the analysis is not to trace influence or reference flows, but to characterise the conceptual landscape surrounding each triad in terms of mutual relevance and clustering, which is precisely what an undirected representation is suited for. The final graph consists of 1017 nodes and 126073 edges.

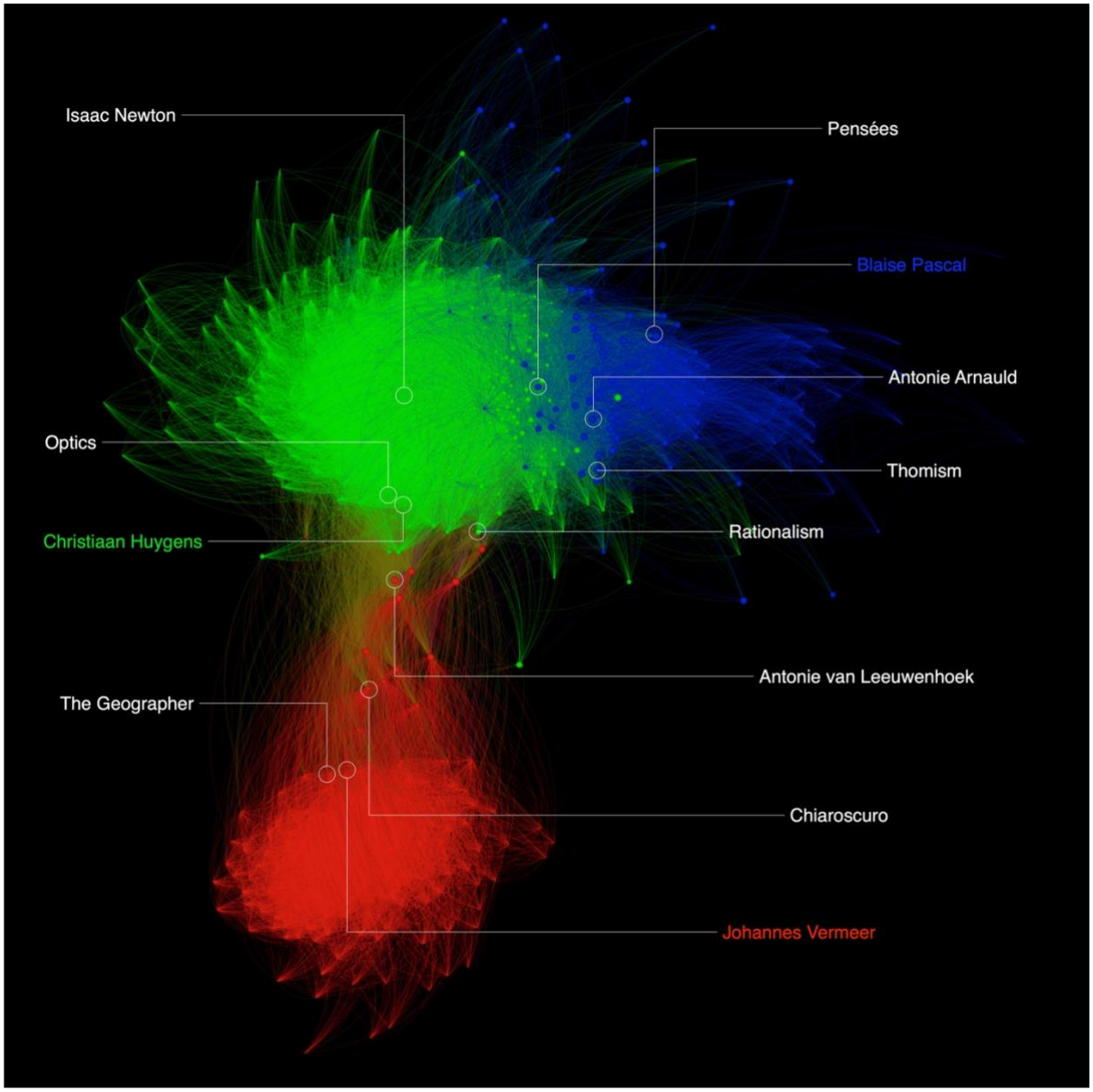


Figure 5 Cultural network based on the seeds of Johannes Vermeer, Christiaan Huygens and Blaise Pascal. Each point represents a Wikipedia entry, and each line represents the relationship between those items. Note the strong interaction between Science (green) and Philosophy (blue) and the relative isolation of Art (red).

The visualisation of this network was done using Gephi [25], an open-source software for exploring and analysing networks, employing the Fruchterman-Reingold method [26] known as the force-directed layout algorithm. This algorithm treats the network as a physical system in which nodes repel each other like charged particles while edges pull connected nodes together like springs. The result is a spatial arrangement in which nodes that are more strongly connected tend to cluster together, and nodes that are more loosely connected drift apart. It is important to note that this spatial arrangement

is determined entirely by the structure of the data: no manual positioning is involved. Each node in the visualisation is assigned to the cluster of whichever seed it is most closely related to, as measured by the NGD. Following the same line of thought, if a node is strongly related to multiple seeds, it is placed in the cluster of the seed to which it is most similar.

By visually inspecting the resulting network for the Vermeer-Huygens-Pascal triad (see **Figure 5**) we see an interesting pattern: the Science cluster (anchored by Huygens) and the Philosophy cluster (anchored by Pascal) are positioned in close proximity to one another, reflecting a high degree of conceptual overlap between these two fields in the 17th century. By contrast, the Art cluster (anchored by Vermeer) is more isolated and less connected to the other two. The algorithm places Isaac Newton firmly within the scientific cluster, positions Pascal midway between Science and Philosophy (a position that makes sense if we consider Pascal's contributions to both mathematics and philosophy of religion) and situates the painterly technique of Chiaroscuro midway between optics and Vermeer's cluster, again a conceptually coherent placement that the algorithm arrives at without any field-specific guidance. This internal coherence provides reassurance that the network construction method is capturing meaningful structure rather than producing arbitrary patterns.

2.2.3. Data Sampling and the Role of Triads

The network presented in Figure 5 is a network built from one triad. Even though, it offers a snapshot of the conceptual relationships among three specific individuals, it cannot by itself characterise the intellectual landscape of an entire historical period. Therefore, to produce a more robust and generalisable picture, the network generation procedure is repeated across all available triads, and the resulting network-level statistics are then averaged. This averaging strategy is analogous to the logic of repeated sampling in statistics: the more data combinations we have the more representative our sample will be.

The number of triads required to achieve a reliable characterisation depends on how heterogeneous the underlying period is, and on the level of precision that the research question demands. However, in any case, the statistical error associated with any averaged quantity decreases as the number of triads increases, following the standard deviation in which the error is proportional to the standard deviation (σ) divided by the square root of the number of sampled triads $\sqrt{N_T}$: $\sigma/\sqrt{N_T}$. In the case of this analysis, the number of triads was maximised rather than estimated from a target precision. In other words, given the pool of people available for the period 1600-1650, all possible combinations of one artist, one scientist, and one philosopher were enumerated, yielding a total of 465 triads. This constitutes the maximum possible sample for the given population of people.

Network Properties and Statistical analysis

Beyond the visual comparison presented in Figure 5, the main analysis (see *Results - section 3.1*) draws on a set of quantitative tools from complex network theory and statistical analysis to examine the structure of the graphs more rigorously. For each graph constructed in the previous step (one per triad) a range of measurements is computed at three levels: the network as a whole (i.e., global properties), groups of nodes within it (i.e., cluster properties), and individual nodes (i.e., node-level properties). The following subsections here describe each measure and the rationale behind its use; the results of this analysis are presented in *section 3.1*.

Global properties

Modularity

One of the questions that arises when examining the cultural network of Science, Art, and Philosophy in the 17th century CE as a whole is whether its nodes tend to organise into distinct, tightly knit groups (usually referred to as clusters or communities) or

whether connections are spread more evenly across the entire network. Modularity is a standard quantitative measure used to answer this type of question.

As shown earlier, the network used in this analysis is an undirected weighted one, which means that the connection (edge) between the nodes does not have a direction and each connection carries a weight which in this case is the result of the NGD as a similarity metric. For this type of network, modularity Q is defined following Newman [27] as:

$$Q = \frac{1}{W}\sum_{c}\left(\frac{W_c}{1} - \frac{S_c^2}{4W}\right)$$

where W denotes the total weight of all links in the network, W_c is the total weight of links that fall entirely within cluster c, and S_c is the sum of weights of all links connected to nodes in cluster c, regardless of whether those links go inside or outside the cluster. The sum is computed across all clusters.

A high modularity value indicates that most of the link weight is concentrated within clusters rather than between them. In other words, nodes within a cluster are more strongly connected to each other than to the rest of the network. In the present study a high modularity score for the Science-Art-Philosophy network—which has been generated by taking into account multiple triads—would indicate that the Science, Art, and Philosophy clusters are each internally cohesive and relatively self-contained: concepts and figures associated with Science would be more strongly related to one another than to concepts and figures related to Art. Consequently, this would suggest that, despite sharing a common historical moment, the three fields maintained distinct conceptual identities in the 17th century. In contrast, a low modularity score would point to a more entangled cultural landscape, in which the boundaries between Science, Art, and Philosophy are blurred and nodes cross cluster boundaries freely.

As mentioned earlier, this part of analysis, is not limited to a single network. More specifically, one network per triad is constructed and the goal is to characterize a

historical period using the averaged metrics from a plethora of triads of individuals from that period. The issue is that each of these networks might differ in size, density, and the distribution of link weights. All these structural differences affect how high the raw modularity score Q can theoretically get, independent of how well-clustered the network actually is. For instance, a network with very unequal cluster sizes, has a lower theoretical ceiling for Q than a network with balanced clusters. This means that if raw Q values are compared directly, modularity could mistakenly show that one triad is less clustered than another, when in reality a triad had simply lower maximum possible score to begin with. For this reason, we start by calculating the maximum achievable modularity for each network:

$$Q_{Max} = 1 - \sum_{c} \frac{S_c^2}{4W^2}$$

Then a normalisation is performed prior to comparison by dividing each network's Q by its own maximum Q_{Max}:

$$Q_N = \frac{Q}{Q_{Max}}$$

This normalisation places every network on the same $[0,1]$ scale where the score reflects how close the network comes to its own best possible clustering and not on some absolute ceiling. This makes cross-triad comparisons meaningful and interpretable.

Cluster properties

Assortativity

While modularity summarises the overall degree of clustering in a single number, the assortativity matrix A provides a more detailed picture by capturing which clusters tend to connect with which. Each element a_{ij} of this matrix records the total weight of links that run between nodes belonging to a cluster i and nodes belonging to cluster j.

Diagonal elements (i.e., where $i = j$) therefore represent internal connections within a cluster, while off-diagonal elements represent inter-cluster connections.

As in modularity, to facilitate interpretation and comparison, the matrix is normalised by dividing each element by the sum of all elements in the matrix:

$$A_N = \frac{A}{||A||}$$

Where $||A||$ denotes the sum of all entries in A. Each entry in the resulting normalised matrix A_N then expresses the proportion of total link weight that flows between a given pair of clusters. This metric can tell us, for instance, whether two particular clusters share a disproportionately large share of connections relative to the rest of the network.

In the context of the present study, the assortativity matrix allows the analysis to go beyond the binary question of whether clusters are well-separated or not, and instead ask a more nuanced question: which fields were conceptually closer to each other in the 17th century, and which were more distant? This is answered by this method, because, for a given triad, the off-diagonal elements of A_N directly quantify the degree of conceptual overlap between pairs of clusters. For instance, a relatively large value in the cell corresponding to the science and philosophy clusters would indicate that substantial proportion of the network's connections flow between those two fields—which would be consistent with the visual observation in the network presented in Figure 5. A small value in the cell corresponding to Art and Science, by contrast, would confirm that the cluster of Art is more isolated from the cluster of Science; something that visualisation suggests too (see **Figure 5**).

This is important for this analysis for two reasons: 1) It provides a quantitative foundation for claims that might otherwise rest solely on visual inspection of the network layout. 2) Because the matrix is computed separately for each triad, it enables systematic comparison across triads. For example, with this method, we can see

whether the relative proximity of Science and Philosophy is a stable pattern that recurs across multiple triads, or whether it is specific to the Vermeer-Huygens-Pascal configuration. In this way, assortativity contributes directly to the broader goal of characterizing the structure of 17th century intellectual culture through the lens of network analysis.

Openness

A complementary perspective on cluster structure is provided by the measure of openness, which shifts attention from the network level to the level of individual nodes within each cluster. A node is considered open if the sum of its link weights to nodes outside its own cluster (its external strength) exceeds the sum of its links weights to nodes within its cluster (its internal strength).

Openness Op for a given cluster is defined as the percentage of nodes in that cluster for which external strength exceeds internal strength. It therefore ranges from 0% (all nodes are more strongly tied to their own cluster than to others) to 100% (all nodes are more strongly connected outward that inward). A cluster with low openness is relatively self-contained and internally cohesive, while a cluster with high openness is more open with its members maintaining stronger ties across cluster boundaries than within them.

In the context of the present study, openness gives an answer to the following question: how self-contained each field was (i.e., Science, Art, Philosophy) in the 17th century? For example, a low openness score for the Art cluster would suggest that the concepts and figures connected to artists of that century were predominantly connected to one another, forming a relatively enclosed conceptual world with limited overlap with Science and Philosophy. By contrast, a high openness score for the Science cluster would indicate that many of the figures and the concepts of scientists of the 17th century were more strongly connected to nodes outside the Science cluster

than within it. A finding like this one would point out to a field that was deeply entangled with neighbouring intellectual traditions rather than operating in isolation.

In historical terms, this is a meaningful difference. If the Philosophy cluster shows high openness, it would suggest that engagement with Art and Science was not limited to a few individuals, but it is a feature that could characterise the entire period under study (the 17th century). On the other hand, if openness is low but assortativity reveals strong ties between specific cluster pairs, it would suggest that cross-field exchange was mediated by a small number of pivotal nodes, which could then be identified and examined more closely.

Node-level properties

Node Connectivity

To enhance the general picture that can be obtained by the previous methodological steps, the focus turns to individual nodes and the specific patterns of their connections. Although the networks studied at this part of the thesis are undirected, it is analytically useful to decompose each node's total connectivity into two components:

- Internal strength (s_{int}): the sum of weighted links connecting a node to other nodes within the same cluster.
- External strength (s_{out}): the sum of weighted links connecting a node to nodes in other clusters.

This decomposition makes it possible to identify, for instance, whether a particular node acts as a bridge between communities and/or a hub within each own cluster. The former is denoted by a high external strength the latter by high internal strength. This information, in turn, provides insight into the structural role that entity plays in the broader network.

In the context of the present study, this measure allows the insights of this study to move from broad structural patterns down to the level of specific concepts, figures, or artefacts and examine the role each of these nodes play in the cultural landscape of the 17th century (see Focus Reading in *Introduction*). This is important because it shifts the focus from the structural to the interpretive. In other words, identifying the nodes with the highest bridging capacity is particularly valuable in a humanities context because those nodes can refer to concepts or figures that played a genuinely cross-field role and now can be studied at a larger scale. Moreover, this kind of finding would be difficult to arrive at through close reading alone, and it illustrates one of the important contributions that a computational study can offer to historical and cultural inquiry.

Bootstrapping

The network measures described above (e.g., Modularity, Assortativity, Openness, and Node Connectivity) are all derived from a single instance of each network. To ensure that the values obtained are stable features of the underlying cultural structure as this structured is presented through Wikipedia at a particular point in time, we use the bootstrapping method.

This method is used in this study to report all the statistical quantities, including means, variances, and derived coefficients. Bootstrapping is a computational resampling technique that estimates the reliability of a statistical result without requiring access to the full population of data [28, 29]. Moreover, it does not assume any particular theoretical distribution for the data. The only assumption behind this methodology is that the sample should be representative of the population which is most likely to happen with a larger number of iterations [30, 31]. The method works by repeatedly drawing random samples with replacement from the dataset, computing the statistic of interest on each resample, and using the distribution of those repeated estimates to characterise uncertainty and construct confidence intervals.

Replacement in the context of the bootstrap method means that the drawn data points in one iteration return to the dataset and can be re-drawn in a next iteration. That means that every observation in the original dataset is eligible to be selected, even if it has already been chosen in that same resample. A single resample may therefore contain some observations multiple times and omit others entirely. The point of this procedure is not to generate new data, but to simulate the variability one would expect if the study were repeated many times under similar conditions. Usually, the resampling process is carried out multiple times (in the case of this study, the resampling was performed 1000 times). After a large number of resamples, the resulting distribution of estimates gives a reliable picture of how stable or variable a given statistic is.

This stability is particularly informative in the study presented here because the analysis compares multiple triads, and it is important to establish that observed differences between them are statistically robust rather than by chance. In general, this approach speaks to one of the central methodological challenges of applying quantitative tools to historical and cultural data. Unlike controlled experimental data, the networks studied here cannot be replicated under different conditions, and the entities they represent (i.e., 17th century scientists, artists, philosophers, and the concepts associated with them) cannot be resampled from some larger population in any literal sense. Bootstrapping offers a principled way to nonetheless reason about uncertainty and reliability within these constraints, lending the quantitative findings a degree of statistical credibility that would otherwise be difficult to establish.

2.3. WikiTextGraph[12]

The methodology introduced in *Section 2.2* retrieved Wikipedia content pages through online API requests, an approach that was well-suited to the bounded task for which it was designed: characterising the cultural network connecting Art, Science, and Philosophy in 17th-century Europe (see *Results - section 3.1*). That study demonstrated that Wikipedia's network can function as a quantitative lens onto cultural history through complex network analysis. However, the individuals under study spanned only fifty years, and the three disciplines examined represented a narrow slice of Wikipedia's full network. These constraints were acceptable at that stage of the analysis, but they become an obstacle when the goal is to extend the temporal window and expand the population of articles to approximately two million biographical pages. At that scale, online API requests are no longer feasible: Wikipedia imposes rate limits on automated queries, and the cumulative overhead of retrieving millions of individual pages through repeated requests makes this approach computationally impractical.

Scaling up this kind of analysis is not only a matter of applying the same pipeline to more data. It introduces two compounding problems that must be addressed together. The first is computational. The English Wikipedia dump runs to several gigabytes in compressed form and can expand to hundreds of gigabytes once unpacked. Loading it into memory on standard research hardware is not feasible, and processing articles one by one, without careful engineering, quickly exhausts available resources and runs for impractically long periods. The second problem is one of accessibility. Researchers without an advanced technical background face an additional obstacle, because working directly with Wikipedia’s raw data requires a high level of programming and systems expertise. Together, these two barriers mean that working with the entirety of Wikipedia’s network may not be as straightforward as it seems and that a software that

[12] WikiTextGraph has been published as a peer-reviewed software tool; the accompanying paper, 'WikiTextGraph: A Python Tool for Parsing Multilingual Wikipedia Text and Graph Extraction,' appeared in the Journal of Open Research Software [32].

will deal with the issues mentioned above is necessary for working with cultural data at scale in the field of digital humanities and cultural analytics.

Existing tools address parts of these problems, but none resolves them in a unified way. General-purpose text extraction tools such as WikiExtractor [33] and Cirrus Extractor [34] are designed primarily to produce clean plain text for Natural Language Processing (NLP) applications; they do not construct link networks. The wiki-dump parser [35] converts Wikipedia dumps into a structured format but relies on WikiExtractor for text cleaning and does not resolve redirects. PyWikibot [36], a Python library maintained by the Wikimedia Foundation, offers real-time data retrieval through Wikipedia's API and handles redirect resolution, but it queries the live website article by article, making it orders of magnitude slower than processing a local dump file directly. Other projects have released pre-built graph datasets for specific languages or time windows [37, 38], but these cannot be easily updated or extended to new language editions or dump snapshots. Taken together, no existing tool combines efficient large-scale parsing, thorough text cleaning, redirect resolution, broken link removal, and graph construction into a single coherent and reproducible workflow that is accessible to researchers without deep programming expertise.

This gap is what motivates the development of a purpose-built tool. The following section therefore introduces WikiTextGraph, an open-source Python library developed as part of this thesis to parse Wikipedia dumps efficiently and to extract the complete internal link network from any language edition of the encyclopaedia [32]. By addressing both the computational and accessibility barriers within a unified workflow, WikiTextGraph makes the large-scale cultural analysis that the remainder of this thesis depends on both feasible and reproducible.

2.3.1. How WikiTextGraph Works

WikiTextGraph[13] takes as input a compressed Wikipedia database file, known as a dump, which the Wikimedia Foundation makes publicly available for every language edition of the encyclopaedia. Compression here means that the file has been algorithmically reduced in size for storage and distribution. These dumps are distributed in XML format, a widely used way of structuring text data so that both humans and computers can read it: content is wrapped in labelled tags that describe what each piece of information represents, allowing the boundaries between articles, titles, and other elements to be identified unambiguously. Together, compression and XML formatting mean that the entirety of a Wikipedia language edition (i.e., all of its articles, titles, and internal structure) is packaged into a single downloadable archive. After this file has been provided to WikiTextGraph, the user also specifies which language edition they are working with and where on their computer the output files should be saved. Optionally, they can request that the tool construct a network graph from the data, in addition to extracting the text. WikiTextGraph operates in two sequential stages: the first is to parse and clean the articles, and the second, is graph construction.

Stage 1: Parsing and Cleaning. Instead of loading the entire dump file into the computer's working memory at once, WikiTextGraph reads and processes the dump incrementally, handling ten thousand articles at a time. By processing articles in fixed-size batches rather than all at once, the tool keeps the amount of data held in working memory at any given moment small, making it possible to run the software on standard research hardware rather than requiring specialist infrastructure.

The parsing method the tool uses is called SAX, which stands for Simple API for XML. Parsing, in this context, means reading a structured file and identifying and extracting the meaningful pieces of information within it — in this case, distinguishing article

[13] https://github.com/PaschalisAg/WikiTextGraph

titles from body text, and both from the many formatting and administrative elements that surround them in the raw dump. Unlike parsing approaches that first read the entire file into memory and build a complete internal map of its contents before any processing begins, SAX reads the file from beginning to end in a single continuous pass, acting on each piece of content as it is encountered and then discarding it. This means the tool never needs to hold more than a small portion of the file in memory at once, making it both fast and memory efficient.

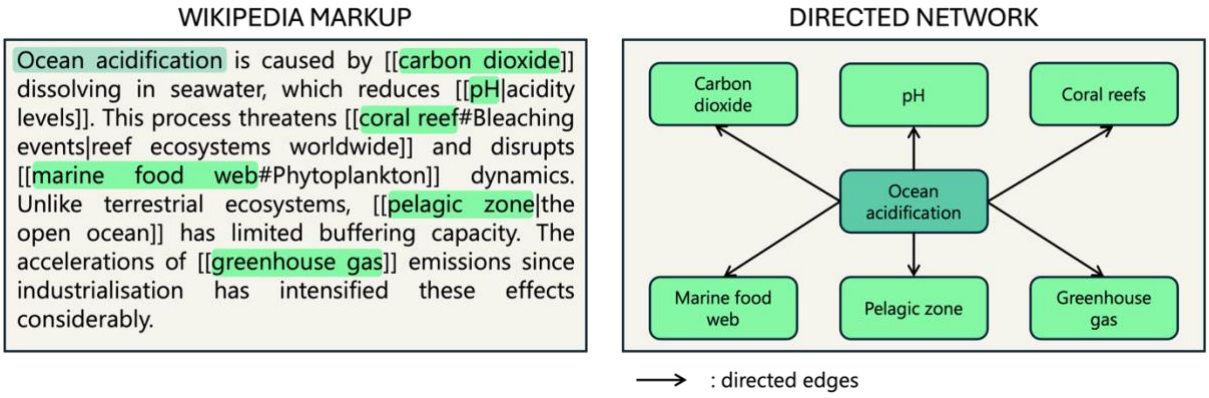


Figure 6 Illustration of the link extraction and graph construction process applied to a single Wikipedia article. The left panel shows a raw Wikipedia article on ocean acidification with wikilinks encoded in MediaWiki markup syntax.[14] Highlighted in green are the six article titles extracted by the regular expression: carbon dioxide, pH, coral reefs, marine food web, pelagic zone, and greenhouse gas. Non-extracted components (e.g., section anchors and display text) remain visually unmarked. The right panel shows the directed network subgraph produced from this article: each extracted title becomes a target node, the source article (ocean acidification) becomes a source node, and each link is recorded as a directed edge pointing from source to target. This figure demonstrates the one-to-one correspondence between link extraction and edge construction that underpins the graph construction pipeline described in this section.

As the tool reads each article, it performs two types of filtering. The first removes entire pages that do not contain encyclopaedic content. Wikipedia contains many pages that serve administrative, navigational, or bibliographical purposes rather than presenting

[14] *MediaWiki markup syntax* is the lightweight markup language used by Wikipedia and other Wikimedia Foundation projects to format article content. It encodes text formatting, structure, and wikilinks using a set of plain-text conventions interpreted by the MediaWiki software at render time. Wikilinks (i.e., links between articles within the same wiki) are written using double square brackets, such that `[[Article title]]` renders as a clickable link to the corresponding article. The syntax also supports display text variants (`[[Article title|display text]]`), section anchors (`[[Article title#Section]]`), and combinations of both (`[[Article title#Section|display text]]`), all of which resolve to the same target article for the purposes of network construction.

knowledge about a subject: templates, which are reusable formatting blueprints that editors apply across multiple pages; category pages, which group articles by theme; redirect pages, which exist solely to forward a reader from one title to another; help pages, which document Wikipedia's own editing conventions; and disambiguation pages, which list multiple articles that share a similar name. These are identified by prefixes in their titles — that is, fixed strings of characters that appear at the very beginning of the page title, such as "Template:", "Category:", or "Help:" — and are excluded from further processing (see *Appendix - A2.1. Page-level filter*).

The second type of filtering removes non-content sections from within articles that do pass the first check. Sections such as "References", "External links", "See also", and "Further reading" appear at the end of most Wikipedia articles and do not form part of the encyclopaedic narrative. The tool stops extracting text from an article when it encounters one of these section headings (see *Appendix - A2.2. Sections filter*).

After these filtering steps, the tool removes additional formatting elements that are part of MediaWiki markup syntax[15] but that are not part of the readable text itself. These include HTML comment tags, which are notes left by editors that are invisible to ordinary readers but present in the raw file; citation reference tags, which encode the footnote markers attached to inline references; and templating constructs such as infoboxes, which are the structured summary tables that appear in the upper-right corner of many Wikipedia articles. Stripping these elements out leaves only the plain prose of each article's main body. The title and cleaned text of each processed article are then written to disk in batches as processing proceeds, rather than being accumulated in working memory until the entire dump has been read. This further reduces the tool's memory demands and ensures that no work is lost if processing is interrupted (for an overview of Stage 1 see **Figure 7**).

[15] For the definition see *Footnote 14*.

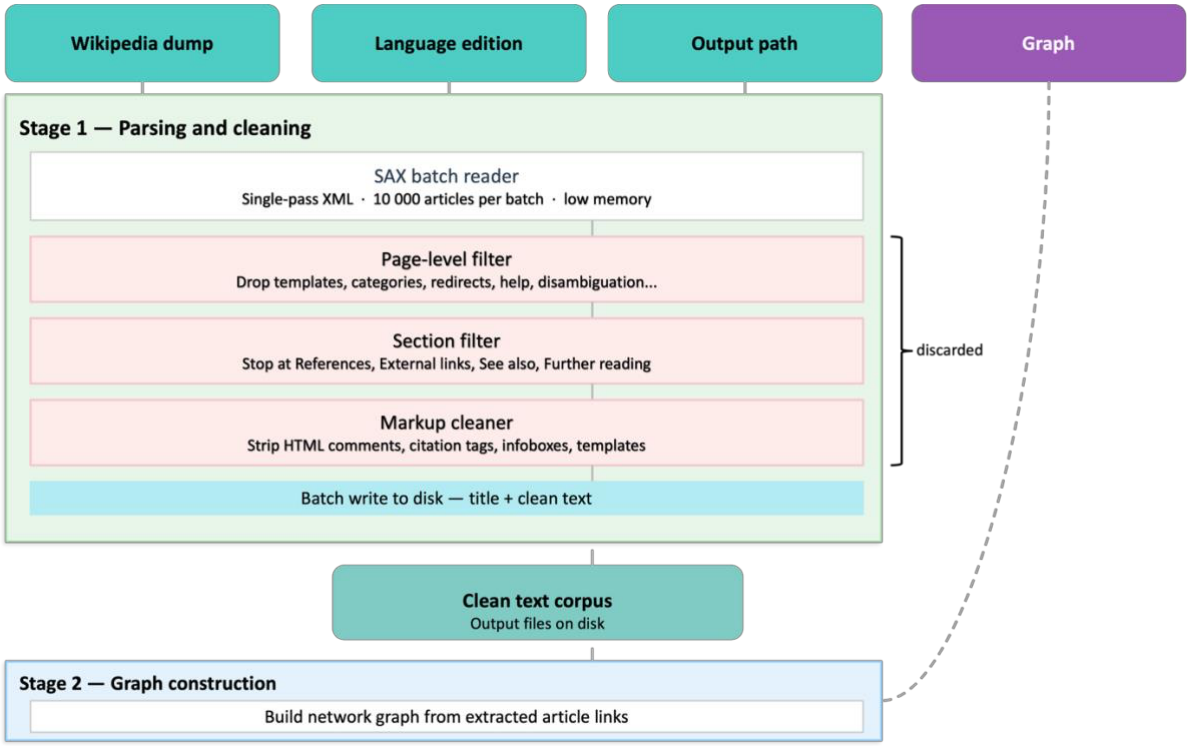


Figure 7 The full pipeline of WikiTextGraph from top to bottom. The teal boxes at the top represent the four inputs. Focusing on Stage 1 the diagram presents how WikiTextGraph, given the input, processes and prepares Wikipedia articles. The output is Wikipedia content articles cleaned and pre-processed. Stage 2 picks up from the clean output and builds the network graph.

Stage 2: Graph Construction. After parsing and cleaning the content articles, WikiTextGraph proceeds to extract the wikilinks from the cleaned articles and uses them to construct a directed network. In this case, each Wikipedia article corresponds to a node, and each wikilink from one article to another corresponds to an edge. The network is directed, meaning that every connection has an explicit direction associated with it: a link from article A to article B is recorded as an edge going from A to B and is treated as distinct from a link travelling in the opposite direction, from B to A.

As mentioned earlier, internal links in Wikipedia are written in a MediaWiki markup syntax using double square brackets (refer to **Figure 6**).[16] The tool uses a pattern-matching procedure, called a regular expression, to identify all such links within each article's text (see **Code Block 1**). A regular expression is a precisely defined sequence of characters that instructs the computer to search a body of text and locate all

[16] Refer to footnote 9 for the definition of MediaWiki markup syntax and a few examples.

passages that match a particular structural pattern, in much the same way that a search function can find every occurrence of a word in a document, but with the added ability to describe flexible patterns rather than fixed strings. This procedure handles the various ways a link can be written in Wikipedia markup, including links that point to a specific section within an article rather than its opening, links that display different text to the reader than the article title they reference, and links that combine both features. In each case, regardless of the specific form the markup takes, the tool extracts only the title of the target article.

Code Block 1 Regular expression used to detect the different types of wikilinks. The regular expression is designed to capture wiki markup patterns (e.g., sections, alternative text, etc.) while preserving only the main article link.

```
r'\[\['

r'([^#|]+)'

r'(?:#([^|]]))?'

r'(?:\|([^]]*))?'

r'\]\]'
```

Before the network is finalised, several cleaning steps are applied to ensure its accuracy and integrity. First, link targets are normalised, meaning they are brought into a consistent form so that equivalent references are treated as identical by the software. Wikipedia article titles can be written with either spaces or underscores between words, and both forms appear in the raw dump. The tool converts all underscores to spaces so that references to the same article are recognised as pointing to the same node regardless of how they happened to be written.

Second, redirects are resolved (see **Figure 8**). A redirect is a Wikipedia page that exists solely to forward readers from one title to another, for example from an alternative spelling, a common abbreviation, or a former name to the canonical article that carries

the full encyclopaedic content. If redirects are left unresolved, what is actually the same article appears as multiple distinct nodes in the network, artificially inflating its size and distorting the connection structure. WikiTextGraph builds a complete mapping of every redirect page to its final target article and replaces all redirect links with the canonical target before the network is assembled. This mapping is stored in a compressed file, meaning it is saved in a space-efficient encoded form, that can also be consulted independently by researchers who wish to study redirect patterns as a phenomenon in their own right.

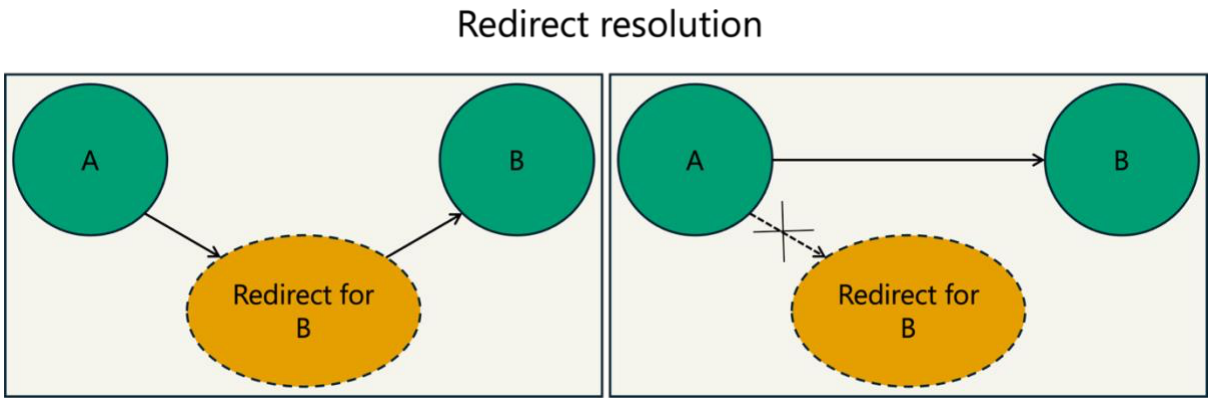


Figure 8 Visualisation of the redirect resolution process. The left panel shows the unresolved state: article A contains a link pointing to a redirect page for article B, which in turn forwards to B itself. The right panel shows the resolved state: WikiTextGraph detects the intermediate redirect, removes the edge to the redirect page, and replaces it with a direct edge from A to B. The redirect node (shown in orange with a dashed border) is excluded from the final network; only canonical article nodes are retained.

Third, broken links are removed. These are links whose target page does not exist in the dump, either because the page was deleted at some point before the dump was created or because it was referenced but never written. Retaining such links would introduce nodes into the network that have no associated article content, which would undermine the integrity of subsequent analyses.

Fourth, links to other language editions of Wikipedia are excluded. Although Wikipedia articles sometimes contain links that point to their counterparts in other languages, these cross-language links are filtered out so that each language edition is treated as

a self-contained network. This ensures that the structure of the resulting graph reflects only the internal knowledge organisation of the edition under study.

Fifth, self-loops and duplicate edges are removed. A self-loop is an edge that connects a node to itself, which in this context would mean an article containing a wikilink to its own page. A duplicate edge is one that records the same connection between the same two articles more than once, which can occur when a link appears multiple times within a single article's text. Neither self-loops nor duplicate edges carry meaningful information for the kind of network analysis this thesis undertakes; thus both are discarded.

The resulting network is stored as a table of source and target pairs, where each row records one directed edge connecting two articles. To allow large graphs to be loaded and processed efficiently, each article is assigned a unique numerical identifier rather than being referred to by its full title throughout the table. Numerical identifiers are faster for computers to compare and look up than text strings of variable length, which is an important consideration when the network contains millions of nodes. A separate lookup table is saved alongside the edge table, mapping each numerical identifier back to its corresponding article title so that results can be interpreted in human-readable terms at any point in the analysis. The edge table itself is saved in the Parquet format, a file format specifically designed for storing large tabular datasets in a way that minimises disk space and allows specific columns to be read into memory without loading the entire file, making it well suited to the scale of data involved here (see **Figure 9**).

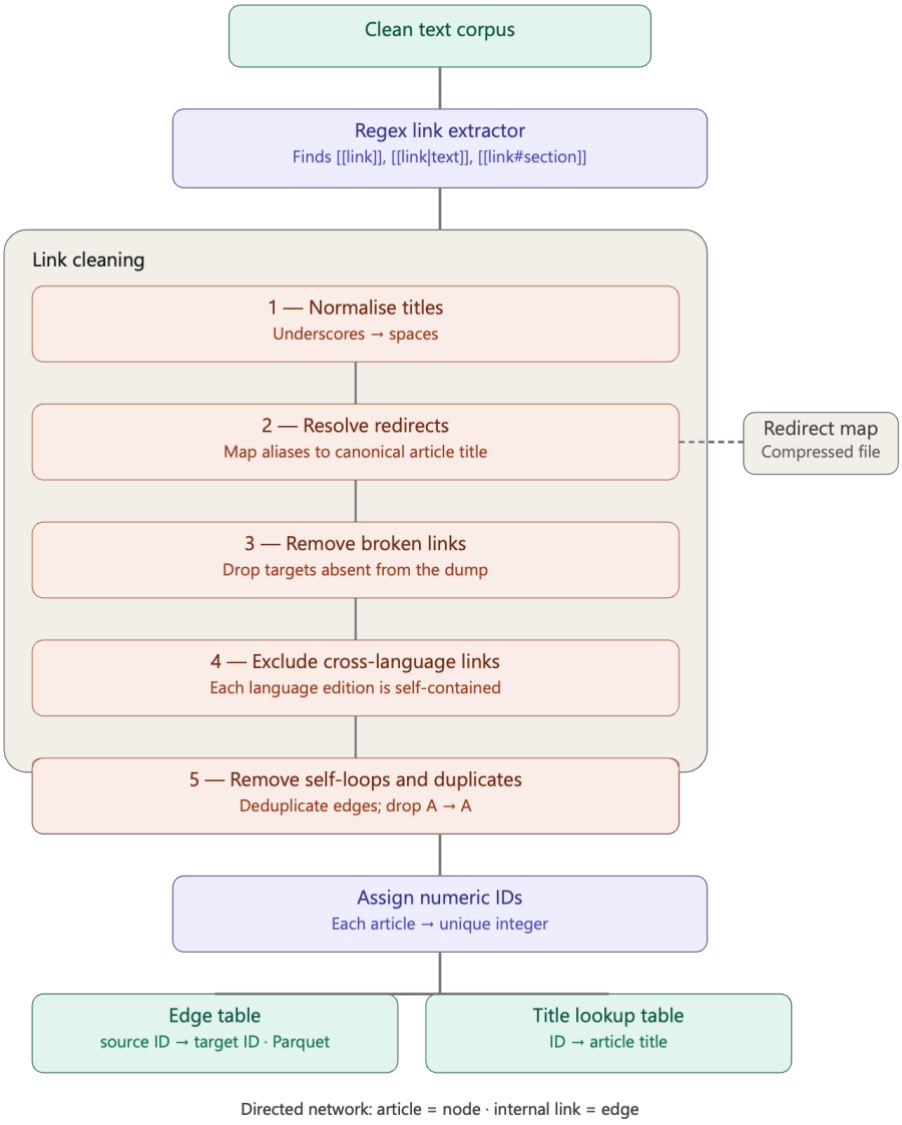


Figure 9 Graph construction pipeline for *WikiTextGraph*. Wikilinks are extracted from the cleaned article corpus via regular expression, retaining only the target article title. Links are then cleaned through five sequential operations: title normalisation, redirect resolution, removal of broken links, exclusion of cross-language links, and deduplication. Each surviving article is assigned a unique integer identifier. The final network is stored as a Parquet edge table of source–target identifier pairs and a title lookup table. Nodes represent articles; directed edges represent wikilinks.

Graphical User Interface (GUI). The tool can be operated in two ways to ensure that users with no technical background can use this software. For instance, users comfortable with the command line can run it by passing the required parameters as arguments to a Python script. Users unfamiliar with command-line environments can instead use a graphical interface, which presents the same options through a point-

and-click window: selecting the dump file, choosing the language edition, specifying the output location, and deciding whether to generate the graph (see **Figure 10**). Once processing begins, the graphical interface closes to free up memory and can be reopened by launching the tool again if needed.

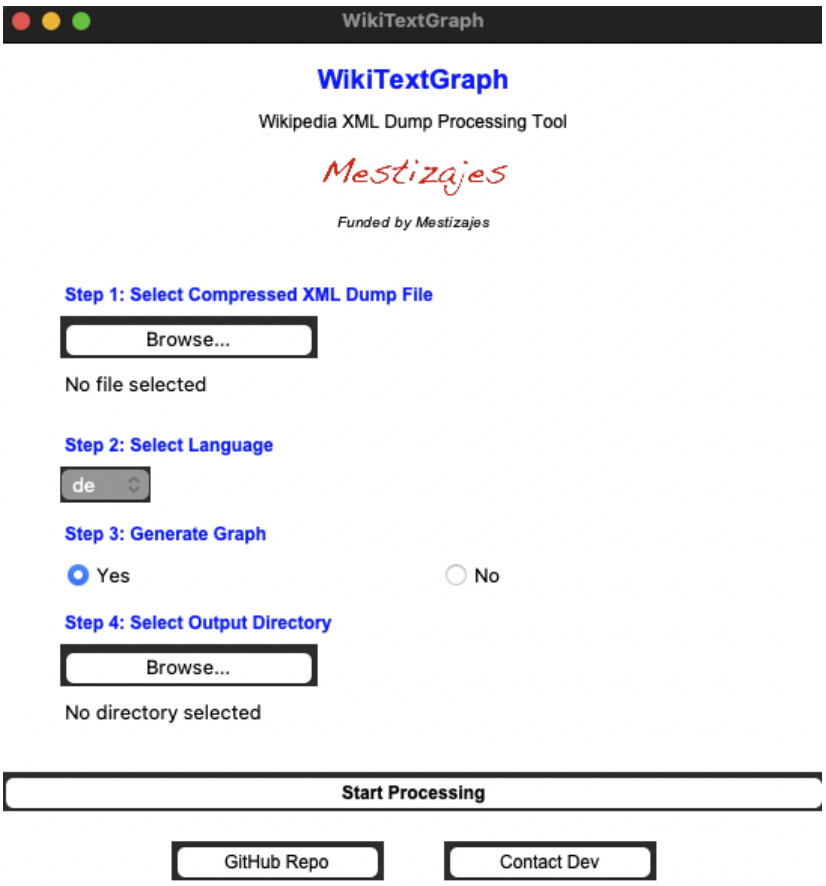


Figure 10 Graphical User Interface (GUI) of *WikiTextGraph*. Users not familiar with running this software through their computer's command line can use the GUI.

2.4. Biographical Network Construction[17]

2.4.1. Introduction

Section 2.2 outlined and explained the methodology used to map the cultural network of Science, Art, and Philosophy within a defined historical window: the first five decades of the 17th century. *Section 2.3* then expanded this scope by introducing WikiTextGraph, an open-source Python library capable of parsing the full English Wikipedia dump, cleaning its articles, and extracting the underlying graph structure. Building on both approaches while addressing their limitations, the present section details the methodology applied to analyse nearly two million biographical entries, with particular focus on the cultural networks of scientists and philosophers across twenty-five centuries.

2.4.2. Extraction and Classification of Biographical Pages

When expanding the scope of the analysis to cover all possible biographical pages on Wikipedia, a fundamental identification problem arises: out of the approximately 7 million articles that make up the English Wikipedia, it is not immediately clear which ones describe real people. Three approaches exist for solving this problem, each with each own limitation.

The first approach, used in *Section 2.2*, relies on the infobox (i.e., the structured summary box that appears in the top-right corner of many Wikipedia articles and contains key facts such as a person's date of birth, nationality, or occupation). While detecting biographical pages through the infobox is a valid strategy [39], many studies have shown that the infobox is not a consistent feature across Wikipedia: a large number of biographical articles either lack an infobox entirely or have one that is

[17] The methodology presented in *Section 2.4* has been presented in the paper "Temporal and Semantic Patterns in Wikipedia's Biographical Network" which has been submitted to the *Journal of Cultural Analytics* and, as of the time of writing (May 2026), it is under review.

incomplete [40, 41]. As a result, any method that relies solely on the infobox will systematically miss a significant portion of biographical pages.

The second approach makes use of Wikipedia's own internal listing system [42, 43]. Wikipedia maintains a set of pages that group articles about people according to various criteria (for example, by occupation, by nationality, or by whether the person is living or deceased).[18] Consulting these lists could in principle yield a more complete set of biographical pages than the infobox method. However, this approach is computationally expensive: it requires fetching and parsing each of these listing pages one by one, then following the links to the individual biographical articles, which makes it impractical when working at the scale of millions of articles.

The third approach, and the one adopted in this study, avoids both of these limitations by turning to Wikidata; a structured knowledge database that runs alongside Wikipedia and stores factual information in machine-readable format [44]. Unlike the infobox method, Wikidata assigns every concept or entity a unique identifier called a "Q-number". It is important to clarify that the Q-number for the concept "human" is Q5, meaning that every person who has a Wikidata entry is explicitly tagged as such, independently of whether their Wikipedia article contains an infobox or appears in any listing page. This makes Q5 a reliable and consistent criterion for identifying biographical pages at scale (see **Figure 11**).

[18] https://en.wikipedia.org/wiki/List_of_lists_of_lists#People

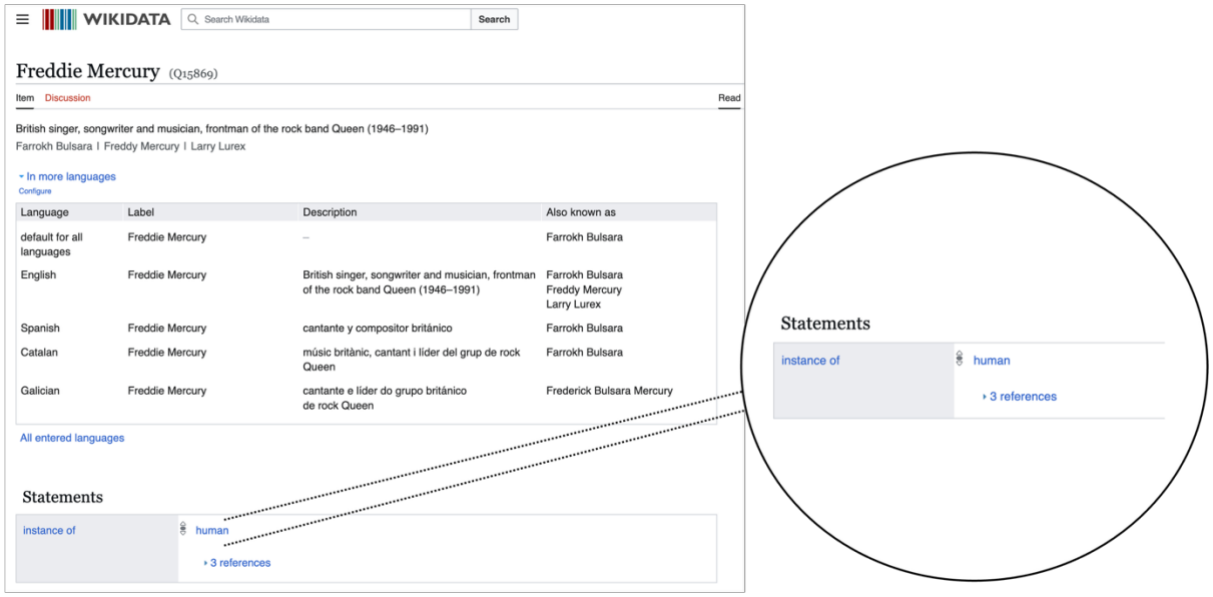


Figure 11 Screenshot of the Wikidata entry for Freddie Mercury (Q15869), illustrating the structure of a Wikidata record. Each entity in Wikidata is assigned a unique identifier (Q-number) and described through a set of structured statements. The magnified panel highlights the "instance of: human" statement, which corresponds to the Q5 identifier, the property used in this study to systematically identify all biographical entities across the English Wikipedia.

To extract all Q5 entities efficiently and avoid the processing delays caused by querying Wikidata's online interface directly — which imposes strict limits on the number of requests that can be made — we instead parsed the full Wikidata dump of January 23rd, 2025. For each Q5 entity found in this dump, we extracted the corresponding English Wikipedia article title, the date of birth, and, where available, the date of death. We then applied WikiTextGraph (see *Section 2.4*) to the English Wikipedia dump of the same date to extract all content pages and the link graph of that version of Wikipedia. By matching the article titles obtained from Wikidata with those extracted from Wikipedia, we obtained a final list of 1,935,343 biographical entries, a number consistent with independent estimates reported by other sources [45, 46] (refer to **Figure 12** for the overview of the step followed).

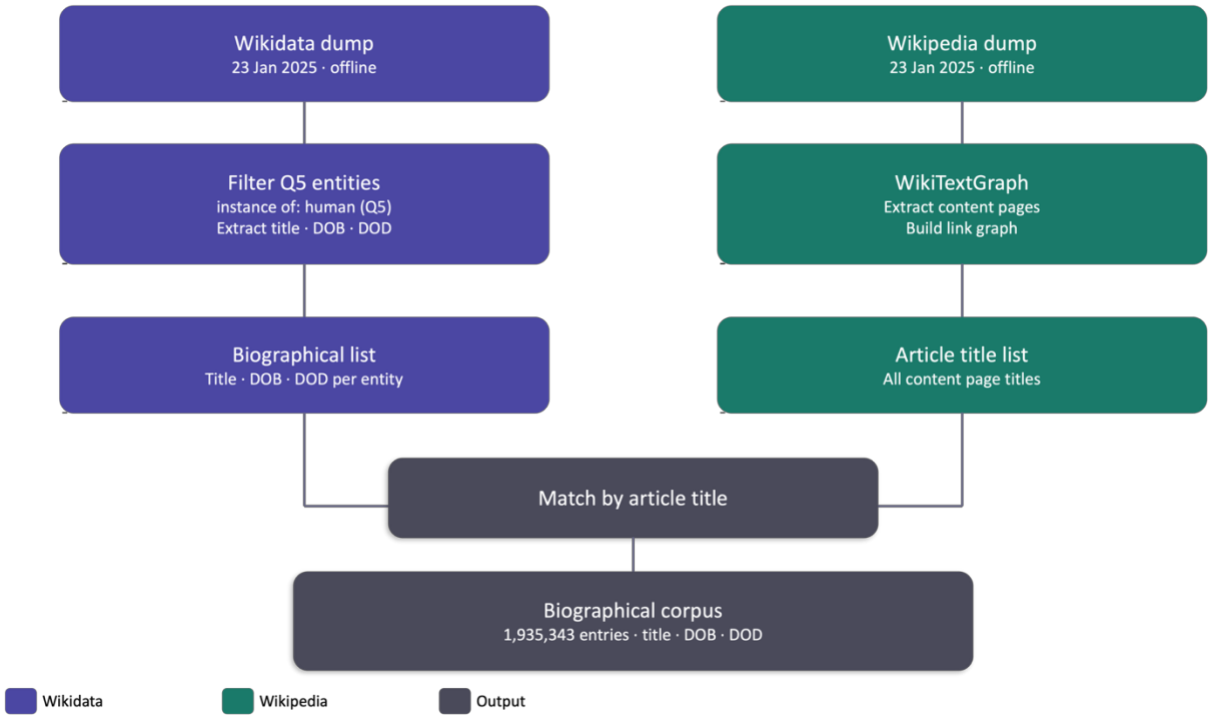


Figure 12 Pipeline for the construction of the biographical corpus. Two parallel processing tracks operate on offline dumps of Wikidata and English Wikipedia, both dated 23 January 2025. The Wikidata track (purple) filters all entities classified as human (Q5) and extracts for each the corresponding English Wikipedia article title, Date Of Birth (DOB), and, where available, Date Of Death (DOD). The Wikipedia track (teal) applies WikiTextGraph to parse all encyclopaedic content pages and construct the internal link graph. The two tracks converge at a title-matching step, which retains only those Wikidata Q5 entities for which a corresponding Wikipedia article exists. The resulting biographical corpus comprises 1,935,343 entries, each associated with an article title, a date of birth, and, where recorded, a date of death.

Once the biographical pages are identified, the next step necessary for the analysis, is to detect what each person is known for. In Wikipedia's editorial conventions, the opening sentence of a biographical article typically contains a brief description of the person's most significant role or contribution. For example, the English Wikipedia page of Grace Hopper presents her notability features as: "an American scientist, mathematician, and United States Navy rear admiral" (see **Figure 13**). Each of these descriptors (e.g., computer scientist, mathematician, United States Navy rear admiral) is in this thesis defined as notability feature. In other word, a notability feature is the occupation, skill, or domain of activity for which a person is primarily recognised on Wikipedia.

Three important things should be noted about the notability features on Wikipedia. First, a single person can have more than one notability feature, as the Grace Hopper example illustrates. Second, notability features are not fixed across languages: the same person may be described differently in the English Wikipedia than in, say, the Spanish or Arabic edition, reflecting different editorial priorities across language communities. Third, and most importantly, notability features represent the judgement of contemporary Wikipedia editors about what a person is best remembered for. For this reason, they do not necessarily reflect how that person identified themselves, or how they were perceived in their own time.

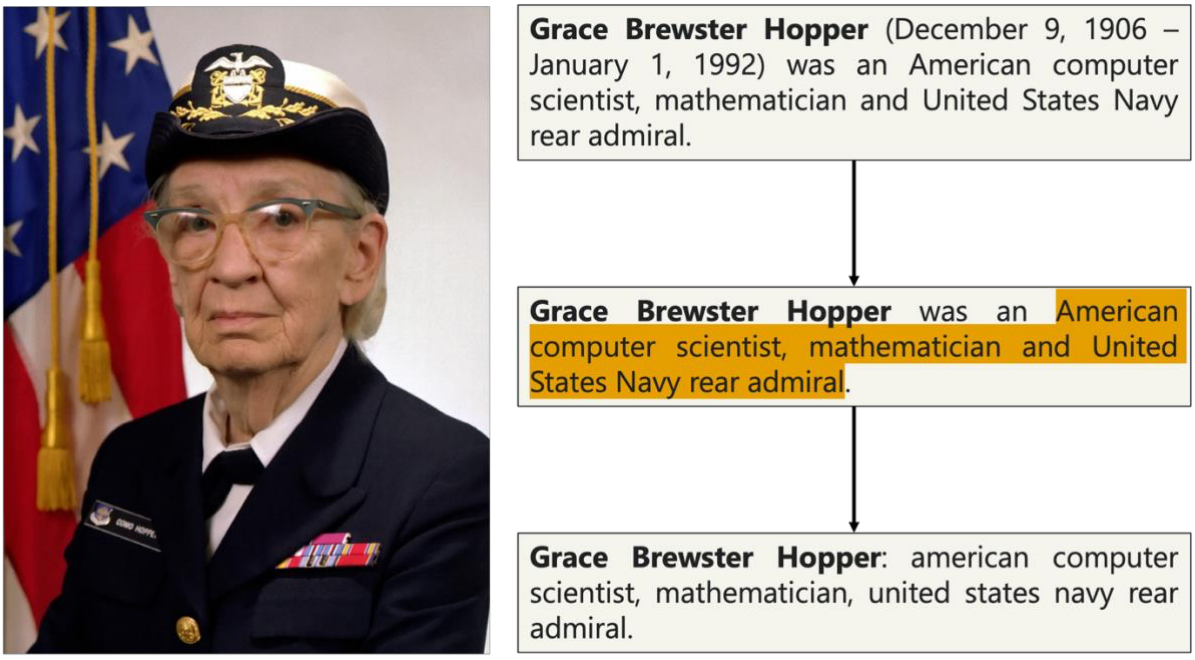


Figure 13 Representation of the detection of notability features using as an example the biographical page of Grace Hopper as of the 6th of May 2026. The upper box shows the first sentence of the biographical article of Grace Hopper as it appears in the English Wikipedia. The middle box shows what the algorithm is going to recognise as notability features from this first sentence. The bottom box shows the final output after cleaning unnecessary text elements from the final output.

Stanza, a software tool that can analyse texts either be breaking it into its components (e.g., tokens) or by assigning part of speech tags to words, was used to isolate the first sentence of each article automatically [47]. More specifically, in the case of this thesis, Stanza was used for sentence tokenisation which is the process of automatically detecting where one sentence ends and the next begins. Once the first sentence was

isolated, a rule-based algorithm was applied to detect and extract the notability features it contains. A rule-based algorithm is a set of explicit, human-written instructions that the computer follows step by step, as opposed to a system that learns patterns from data. The algorithm operates in four stages:

First, the sentence is cleaned: leading and trailing spaces are removed, and any parenthetical content — such as dates of birth, pronunciation guides, or Wikipedia formatting templates — is stripped out, since these elements could otherwise interfere with the extraction. The text is also lowercased to ensure consistent matching. Second, the algorithm searches the cleaned sentence for the first occurrence of a copular or function verb — that is, a linking verb such as is, was, were, served, or played — which in biographical writing typically introduces the description of who the person is. Everything that follows this verb is treated as the region where the notability features are most likely to appear. If no such verb is found, the biography is skipped, as no features can be reliably extracted. Third, the text in this candidate region is broken into individual tokens, and the algorithm reads through them one by one until it encounters a clause-boundary word — such as who, which, for, from, or best — that typically signals the start of additional context rather than a core description (for example, "who was born in..." or "best known for..."). Everything before this boundary word is kept as the candidate descriptor. Fourth, since a person can have more than one notability feature, the candidate descriptor is split wherever a comma, semicolon, or coordinating conjunction such as and appears — reflecting the common Wikipedia pattern of listing multiple roles in sequence, as in "scientist, mathematician, and rear admiral." Two fixed expressions, rock and roll and track and field, are deliberately protected from this splitting step, as their internal and is part of the phrase itself rather than a list separator. The resulting segments are then cleaned of leading articles (a, an, the) and punctuation, yielding the final list of notability features for each biography.

Having the notability features extracted, the next step is to organize them into a set of categories. The previous approach used in *Section 2.2.2* used a limited set of keywords

to perform that task. Even though that approach worked well for a limited number of biographies, in the case of this analysis the heterogeneity of fields in which an individual can belong raises sharply, thus the approach should be broad enough to be able to recognise fields that are not so common.

Previous studies mostly have used the categories under Wikipedia native lists to classify individuals into their fields of activity [12, 48]. However, as some studies have shown, Wikipedia's native category system is unstable and complex which makes the classification task even more difficult [48, 49]. At the same time, as in the case of infobox to detect biographies, nothing ensures that all biographies have been listed in one of Wikipedia's native category systems. For these reasons the approach followed here uses embeddings as a classification task.

The classification task was carried out through a three-step computational process. First, the text was standardised to ensure consistent orthography across the same terms. For instance, each word is lowercased and lemmatized which means that each word is reduced to its dictionary form (e.g., "writers" becomes "writer"). This ensures that slight spelling or grammatical variations do not prevent two phrases that mean the same thing from being recorded as equivalent.

Second, each text is converted into numbers (embeddings). In other words, each standardised notability phrase is converted into a numerical vector using a machine learning model called Sentence Transformer [50].[19] A vector is simply an ordered list of number (in this case, 768 numbers) that encode the meaning of a phrase in a form a computer can compare and calculate with [51]. One feature of these vectors that should be remembered is that phrases with similar meanings end up close to one another in this numerical space. This type of embedding (i.e., Sentence Transformer) is

[19] The model used is https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

chosen because previous studies have shown that contextual embeddings outperform static embeddings [52–54].

Third, once all notability features and Fields of Human Activity (henceforward: FHAs)[20] are represented as vectors, agglomerative clustering is performed which is a bottom-up grouping process that starts by treating every phrase as its own group and then progressively merged the most similar ones. It should be noted that two phrases are considered similar if their vectors are close in the numerical space. How close two vectors are to each other is measured using cosine similarity which measures the angle between two vectors: a value of 1 means that the two vectors are exactly the same, and a value of 0 means that two vectors are completely unrelated. For example, "Spanish author" and "French author" will be merged into the same cluster because their meanings are similar but not exactly the same. For each resulting cluster, mean pooling is applied which results in a representative vector after averaging all vectors in it. This representative vector is then compared to the pre-defined vectors for each of the 13 FHAs: Sports, Government & Law, Dramaturgical Arts, Applied Arts, Music, Literature, Humanities & Philosophy, Business, Science, Religion, Media, Military, and Healthcare & Medicine. Whichever FHA vector was closest determined the category assigned to that cluster. So, for example, a cluster centred around the concept of "athlete" would be matched to the FHA of Sports.

We chose to group them into 13 broad FHAs (see *Appendix A – A3.1. Fields of Human Activity (FHAs)*). The final classification scheme was refined through an iterative process: starting from previous studies [48] and continuing with manual validation of the results to ensure that the majority of the notability features fit into the defined FHAs. Furthermore, to evaluate the reliability of this assignment process, a random sample of 500 biographies is manually annotated, and these annotations were

[20] In this analysis we use the term "Field of Human Activity" as proposed by Eom and Shepelyansky [55] because I believe that it describes more clearly this types of classes. However, it is worth mentioning that previous studies have used terms like: *categories* [48], *cycles* [9], and *occupations* [56].

compared to the algorithmic assignments. Performance is evaluated using normalized accuracy (0.978) and multi-label precision (0.983), recall (0.986), and F1 score (0.985).

At this point, a note on retrospective labelling is necessary. Assigning a label such as "Science" to a figure from Classical Antiquity or the Middle Ages is, strictly speaking, an anachronism: the concept of Science as a distinct discipline did not exist in those periods. A scholar like Archimedes, for instance, would not have identified himself as a scientist or an engineer — these are modern categories. In his own time, his work would have been understood as part of a broader, undivided intellectual tradition that we might today call Natural Philosophy, in which mathematics, the study of nature, and philosophical reasoning were not yet separated into distinct fields. Similarly, many medieval scholars were polymaths (i.e., individuals who worked across what we would now consider multiple disciplines simultaneously) making it impossible to assign them cleanly to a single modern category.

Despite this, the FHA labels used in this study are not applied arbitrarily. Because Wikipedia is built and maintained by contemporary editors, its biographical articles already describe historical figures using modern terminology. Archimedes, for example, is described on his English Wikipedia page as a mathematician, physicist, engineer, and inventor — all modern labels applied retrospectively. This means that the notability features extracted from Wikipedia's first sentences already carry this modern framing, and the FHA labels assigned in this study reflect that framing rather than imposing an additional layer of anachronism.

The use of 13 broad FHA categories is therefore a deliberate methodological choice, made for two practical reasons. First, finer-grained subcategories (e.g., "astrophysics" or "organic chemistry") would be historically meaningless for the majority of the twenty-five centuries covered by this study and would produce extremely sparse data for earlier periods. Second, broader categories provide a stable and consistent framework across the entire time span, making long-term comparisons possible.

2.4.3. Biography-to-biography network

At this stage of the methodology, three pieces of information are available for each person in the dataset: their English Wikipedia title, the century in which they were born, and the FHAs they have been assigned to in the previous step (see *section 2.4.2*). The next step is to create a biography-to-biography network, which is a structured map of the wikilinks that connect biographical Wikipedia articles to one another.

To understand what this means, it helps to recall what a network is in the most basic sense: a collection of nodes (the things being connected) and edges (the connections between them). In this network, each node is a Wikipedia biographical article — representing one person — and each edge is a directed wikilink from one biographical article to another. As mentioned earlier, the word "directed" here is important because it means that a link has a clear origin and a clear destination. For instance, as of the time of writing, the English Wikipedia article on Aristotle contains a wikilink to the article on Galileo Galilei but not the other way around.

To construct the biography-biography network, I begin with the complete English Wikipedia network of content pages and then filter it to retain only biographical pages as sources and targets. This filtering step produces a subnetwork that is substantially smaller than the full Wikipedia graph and qualitatively different from it: every node in the resulting network represents a person, and every directed edge is a hyperlink from one biographical article to another. The resulting directed network contains 1,360,200 nodes and 7,998,710 edges.

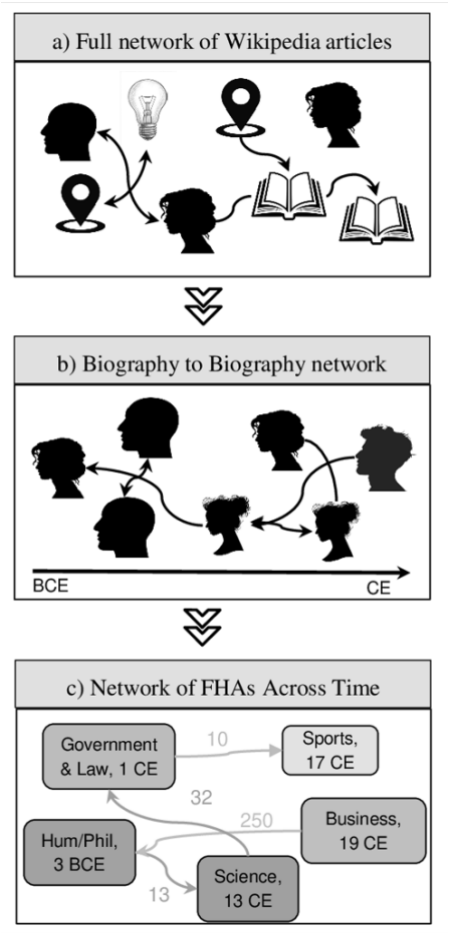


Figure 14 Overview of the data processing pipeline, from the entire Wikipedia network of content articles to the network of notability features over time. (A) The full network of English Wikipedia articles, where nodes represent individual pages and edges represent wikilinks. (B) The biography-to-biography network, created by filtering the full network to include only links between biographical articles. (C) The network of notability features over time, generated by grouping biographies by FHA and century of birth. This representation allows us to analyse how different FHAs are connected across historical periods, as reflected in Wikipedia's internal link structure.

At this point, it is worth pausing to note what a link in this network represents. When a Wikipedia editor writes an article about a person and inserts a wikilink to another person's article, that is a deliberate editorial choice; a statement, however implicit, that these two individuals are related or relevant to one another. The biography-to-biography network therefore captures not just who is mentioned alongside whom, but also the retrospective judgements that contemporary Wikipedia editors have made about which historical figures belong in each other's company (see *Introduction*).

The goal of this analysis is to focus on the broad historical patterns that emerge from the connections between people of different fields. However, working directly with

nearly 1.4 million nodes and almost 8 million edges means that there is a level of detail that is too fine-grained to reveal the broad structural patterns that emerge across FHAs and centuries. Therefore, to make the data manageable and to enable comparisons across long periods of time, the individual biographical nodes are grouped together into their FHAs, producing what is called here the Network of FHAs Across Centuries.

The grouping works as follows: Every biographical page is placed into a group defined by two properties simultaneously: their FHA and their century of birth. For example, all scientists born in the 17th century form one group, all scientists born in the 18th century form another, all philosophers born in the 17th century form yet another, and so on. Each of these groups becomes a single node in the new, aggregated network. A node might therefore be labelled "Science, 17th century" or "Humanities & Philosophy, 5th century BCE."

Once the nodes are defined in this way, the edges between them are constructed by counting how many wikilinks connect the biographical articles in one group to those in another. If, for example, the Wikipedia articles of 17th-century scientists collectively contain 450 wikilinks pointing to articles about ancient Greek scientists, then the edge from the node "Science, 17th century" to the node "Science, 5th century BCE" carries a weight of 450. Weight here simply means the strength or intensity of the connection. In other words, a “heavier” edge means more links flow between those two groups. The result is a weighted directed network, where both the direction of the connections and their numerical strength are preserved (see **Figure 14**). This aggregation step is what makes large-scale historical patterns to emerge. Instead of asking "does Newton link to Aristotle?", the network now allows us to ask questions at a much broader level: "do scientists of the Modern Era tend to link to scientists of the Classical Era, and if so, how strongly?”.

2.4.4. Weight Normalisation & Network Filtering

The previous step constructed a direct weighted network of biographies. However, the weight of edges is the raw count of links which raises an issue when comparing individuals from different historical periods. Some centuries are represented by thousands of biographies, while others in the distant past contain only a small number. Additionally, a group with more outgoing links will produce more outgoing links simply because there are more articles to link from and not necessarily because this group is historically more connected. At the same time, more recent centuries tend to have far more Wikipedia coverage than earlier ones, which could make the Modern Era appear artificially dominant in the network. To address this issue, the raw link counts are normalised (i.e., converted from absolute numbers into proportions) so that groups of different sizes can be more fairly compared with each other.

This normalization is performed in two complementary ways, each answering a slightly different question. The first is the out-degree normalization, which looks at the network from the perspective of the source (i.e., the group that sends the links). For each pair of nodes i (source) and j (target), the raw number of connections w_{ij} is divided by the total number of links that node i sends out to all other nodes[21]:

$$w_{ij}^{+} = \frac{w_{ij}}{\sum_{j=1}^{N} w_{ij}}$$

The result w_{ij}^{+} can be read as: "Out of all the links that source group i sends out, what fraction goes to target group j? In more simple terms, this tells us about the linking priorities of the source group; where the "attention" of group is directed (for an example see *Appendix A – Weight Normalisation – Out-degree normalisation*).

The second approach is the in-degree normalization, which approaches the networks from the perspective of the target group; the group that receives the links. Here, the

[21] A node, as mentioned earlier, is a combination of the FHA and the century.

raw number of connections w_{ij} is divided by the total number of links that node j receives from all other nodes:

$$\bar{w_{ij}} = \frac{w_{ij}}{\sum_{i=1}^{N} w_{ij}}$$

The result $\bar{w_{ij}}$ can be read as: "out of all the links that the target group j receives, what fraction comes from the source group i? This tells us about the historical visibility of the target group. In other words, which groups reference it, and how much (for an example see *Appendix A – Weight normalisation – In-degree normalisation*). The reason both perspectives are important is that they can answer different questions from the same underlying data. Moreover, they can reveal some asymmetries that might have been encoded in the data that would not be visible if only one of the normalisations was considered. Lastly, as the results show (see *Results – Section 3.2*) the choice of normalisation is not merely a technical detail, but it affects the historical narratives that emerge from the data.

Once the weights have been normalised, a three-step filtering process is applied before the main analyses. The purpose of this filtering is to remove parts of the network that are too sparse or too unbalanced to produce reliable results. Groups with very few biographies or very few links could distort the patterns observed, since a single unusual link would have an outsized influence on a group that contains only two or three people. The three filters are applied sequentially, and the entire sequence is repeated in a loop until the network stabilises and no further nodes are removed. Specifically: Filter 1 removes any individual biographical page that has five or fewer outgoing links, since pages with very few links provide insufficient information about a person's connections; Filter 2 removes any FHA–century node that contains five or fewer biographies, since a group that small cannot be considered representative of a broader field or period; and Filter 3 removes any century that is represented by fewer than five different FHAs, ensuring that each century retained in the dataset reflects a sufficiently diverse range of human activity to support meaningful comparison.

It is important to acknowledge the limitations of this filtering process. Some historically significant figures are excluded because their century is too sparsely documented in Wikipedia to meet these thresholds. Confucius and Sun Tzu, for example, do not survive the filtering process, not because they are unimportant, but because the 5th century BCE is represented by too few biographies across too few fields to be retained. This is a limitation of the approach, but it is a necessary compromise: retaining poorly documented centuries would introduce noise and statistical unreliability that would undermine the validity of the comparisons made across the remaining twenty-five centuries. Moreover, the threshold of five used across all three filters may appear arbitrary, but it reflects a deliberate decision to balance between two competing risks: setting it too low would allow centuries and fields with negligible data to distort the results, while setting it too high would exclude a disproportionate number of biographies and reduce the historical coverage of the dataset. After the filtering converges, the dataset spans biographical articles from the 5th century BCE to the 20th century CE, with fields such as Science, Humanities & Philosophy, Literature, Government & Law, Military, and Religion represented across most of this range (for the final result see **Figure 15**). This aggregation yields a weighted directed network in which each node corresponds to a unique combination of century and FHA, and each directed edge indicates that at least one biographical link exists between a source node and a target node, whether within the same century or FHA, or across different ones. The network contains 233 nodes and 18,097 edges. Each edge carries a weight that reflects the normalised number of connections between the corresponding century-FHA pairs, calculated using the normalisation procedure described above.

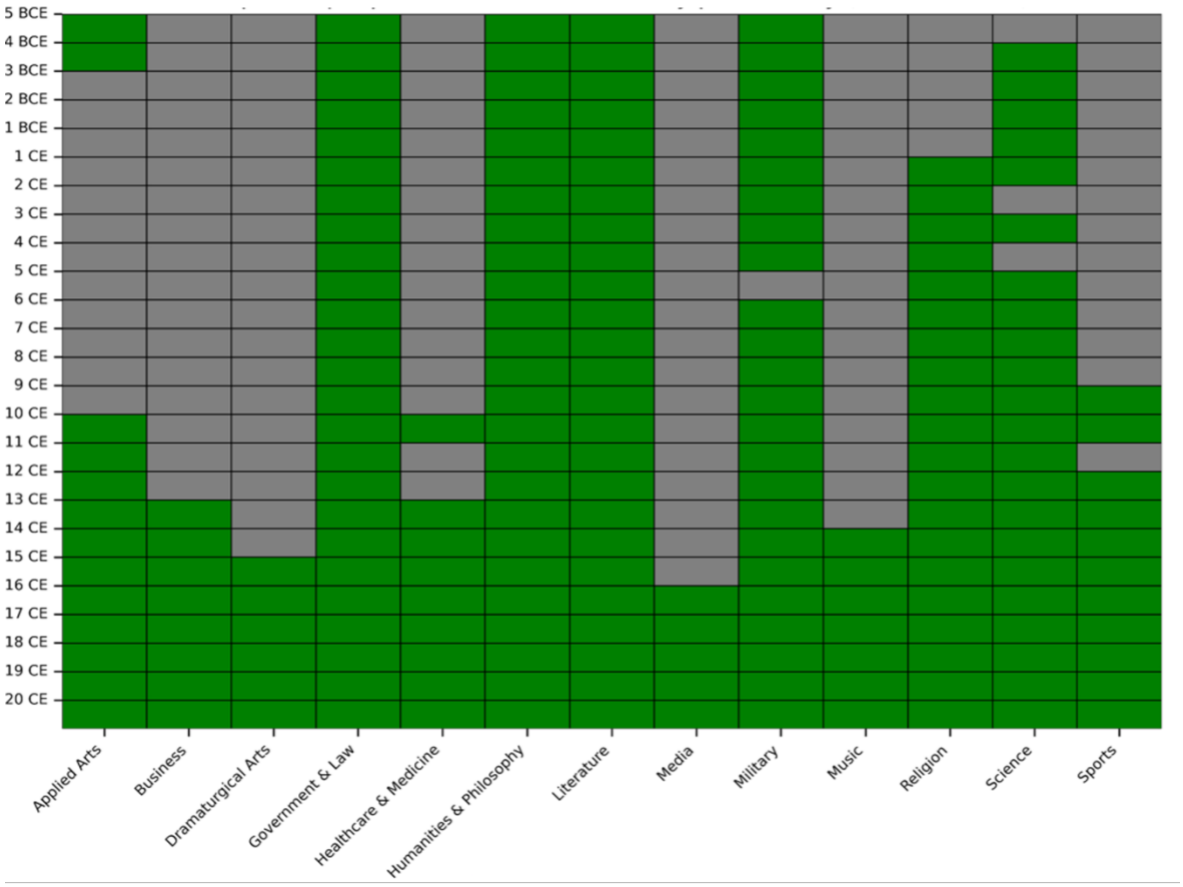


Figure 15 Heatmap of the final iteration. Green cells represent the nodes (defined by century and FHA) that are retained in the dataset after the data filtering control process.

2.5. Semantic Analysis

The methodology described in *section 2.4* constructs a network where nodes are biographical pages and edges are the wikilinks between them. Grouping biographies into their respective FHAs makes it possible to study how interactions between fields such as Science and Humanities & Philosophy changed across centuries, and to identify periods of increased intellectual exchange. But this approach is limited to the structure of connections. It can tell us which fields link to which, and how often. It cannot say anything about the context in which those links occur. For instance, grouping scientists by century of birth might reveal that the concepts appearing in their linked biographies shift gradually over time. This could mean that scientists are anchored to concepts associated with people from earlier centuries, which then give

way to newer ones. Network structure alone cannot answer that kind of question. To address this, the final part of the analysis (see *Results – Section 3.2*) turns to contextual embeddings. These draw on advances in NLP, where contextual language models encode words and entities as vectors that capture semantic relationships based on usage. Within a digital humanities and cultural analytics framework, embeddings make it possible to analyse cultural and historical meaning at scale. They complement network structure with interpretable representations of how concepts and people are situated within evolving textual contexts.

2.5.1. Embeddings

As very briefly explained in *section 2.4.2*, an embedding is a way of representing the meaning of a piece of text as a vector. A vector of numbers is the component that a computer needs to understand the difference between the words "physicist" and "soldier". For this reason, to perform any kind of meaningful comparison between pieces of text, that text must first be converted into a numerical form that a computer can work with. What makes embeddings extremely useful in a computational analysis of textual data is the conversion from alphanumeric characters to a vector because this process is designed to preserve meaning and be "understood" by a machine. Two phrases that are semantically similar (e.g., refer to related concepts or are used in similar contexts) will be mapped to vectors that are numerically close to one another in this space [57, 58] (see **Figure 16** for a schematic illustration).

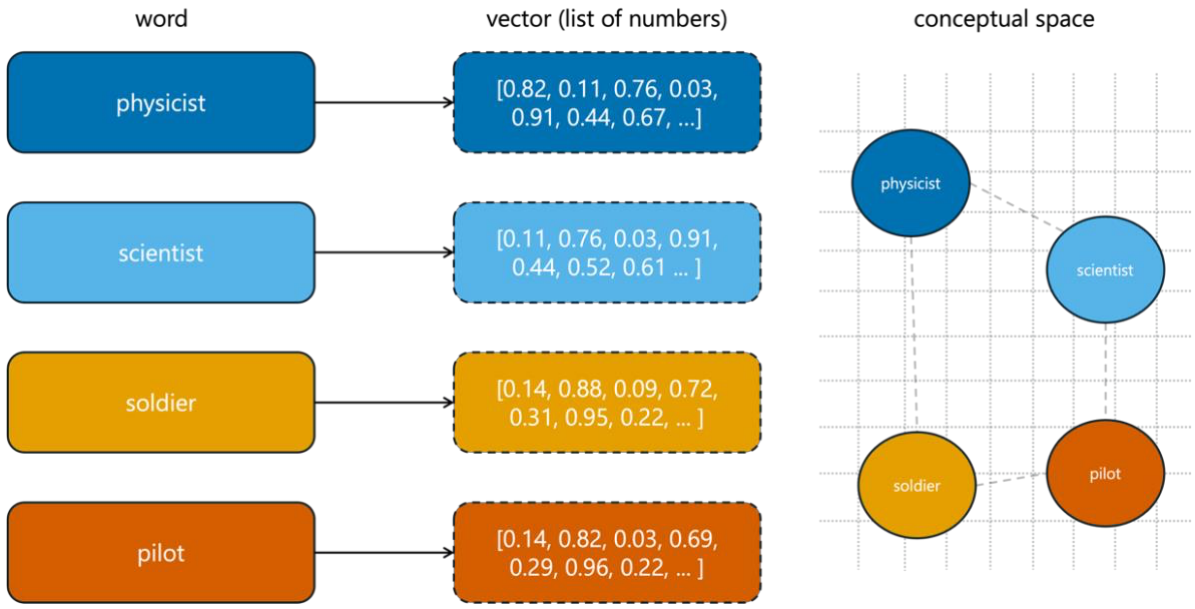


Figure 16 Schematic illustration of the embedding process applied to four notability features. Each word (left column) is converted by the embedding model into a vector; its vector representation (centre column). These vectors are then placed as points in a conceptual space (right panel), where the distance between any two points reflects the degree of semantic difference between the corresponding words.

This way of representing meaning aligns this type of analysis with the tradition of cultural cartography, which aims to simplify texts in faithful ways while preserving the graded, relational meanings of words [58–60]. Embedding models offer a concrete means to do this: they allow the comparison of frequencies to be substituted with the comparison of distances, by providing standardised maps of meaning space in which semantic relations are realized as "spatial" relations. This is a very important aspect for this study, since the questions of interest are inherently graded rather than discrete; the network approach presented earlier to detect hotspots of connections between FHAs follows also a graded approach. It is important to note, though, that embedding models do not "understand" meaning in substantive sense, nor do the procedures built on top of them substitute for interpretation [61, 62]. Rather, it would be more accurate to talk about condensing information in order to facilitate interpretation [63].

The specific type of embeddings used here are known as contextual embeddings, produced by a Sentence-Transformer model.[22] Unlike static embeddings which assign a single fixed numerical representation to each word regardless of context, contextual embeddings capture meaning as it functions within a given sentence or passage [64]. For instance, the word "bank" would receive different vector representations depending on whether it appears in a text about finance or about rivers. Since this part of the analysis aims to study how the context of outgoing links changes across time given an FHA, this sensitivity to context makes contextual embeddings better suited to the kind of nuanced semantic analysis required by the present study. Moreover, in many cases they have been shown to outperform static embeddings in comparable tasks (e.g., classification) [52–54, 65].

In practice, contextual embeddings are used in this study as follows: For every biographical article in the English Wikipedia corpus, all outgoing wikilinks are extracted. Each of these link titles is then passed through the Sentence Transformer model mentioned above, which converts it into a 768-dimensional vector. That number refers to the number of coordinates used to describe the item's position in the semantic space.

Once all the outgoing wikilinks from biographies belonging to a given century are embedded in this way, their individual vectors are averaged together using a technique called mean pooling (see **Figure 17** for an overview of the approach):

$$\vec{v}_c = \frac{1}{n_c}\Sigma_{i=1}^{n_c}\vec{v}_i.$$

This averaged vector $\vec{v}_c$ corresponds to the centroid of the link embeddings. After computing the centroid, L-2 normalisation is applied to ensure that the resulting vector

[22] As in *section 2.4.2* the model used is: https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2.

has unit length. In other words, that the resulting vector represents direction only, without allowing the magnitude to influence the result:

$$\vec{v}'_c = \frac{v_c}{||v_c||_2}.$$

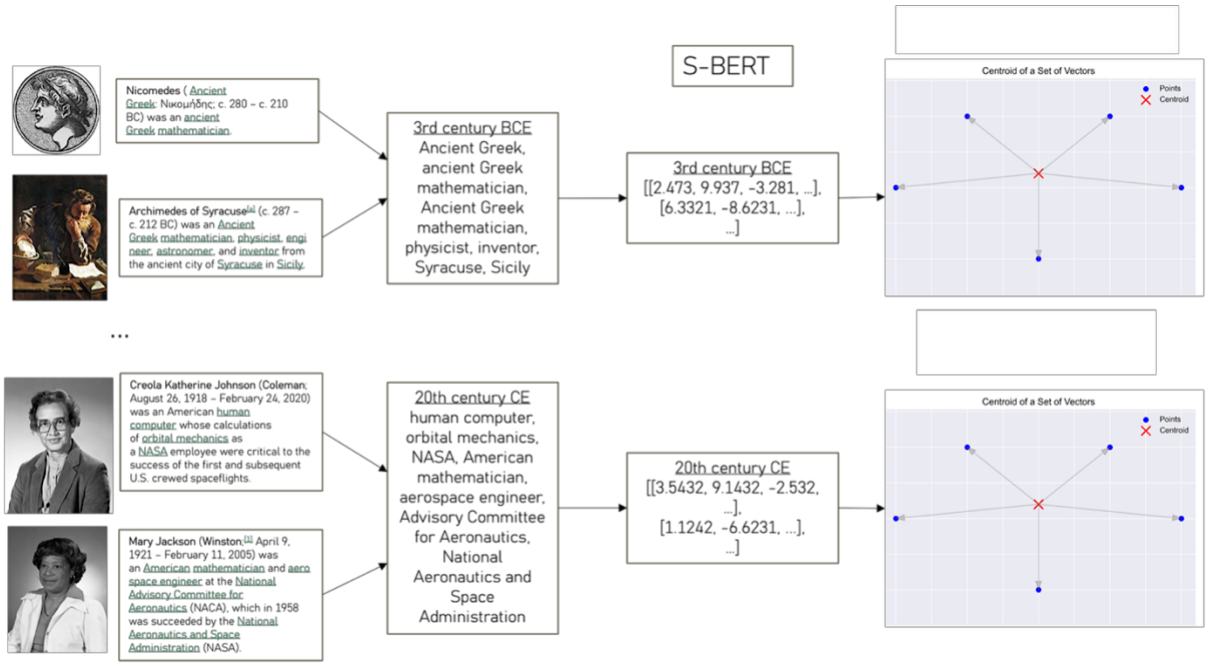


Figure 17 Computational workflow for constructing century-level semantic profiles. The method involves extracting all outbound links from biographies of a given FHA and grouping them by the subject's birth century. These links—covering various entities such as individuals, concepts, and objects—are mapped into a high-dimensional embedding space. Using mean pooling, a centroid vector is computed for each century to represent the central trend of the period's information landscape.

Once each century is represented by a normalised semantic profile, pairs of centuries can be compared using cosine distance, a standard measure for embedding similarity [64]. Cosine distance captures how similar the directions of two vectors are which means that if vectors point in the same direction, they have distance 0, while increasingly different directions yield larger distances. Because the semantic profiles of the centuries are L2-normalised, in my case, vector length does not affect the comparison [64]:

$$d_{cos}(v, w) = 1 - v \cdot w$$

In the context of the analysis presented in *Section 3.2*, the semantic profile computed for each century can be read as a compact summary of the aggregate referential field

invoked by Wikipedia biographies of people born in that century, as encoded by a particular Sentence-Transformer. Taken together, these profiles can be understood as pointing toward something broader which we call here the semantic cultural landscape of a given period. This is the wider terrain of references, associations, and conceptual neighbourhoods within which the biographies of that century are situated. However, there are some points worth noting. First, the profile is corpus-relative: it reflects the references made by present-day English Wikipedia editors rather than historical reality, and a different language edition would yield a different profile for the same century (see *Introduction*). Second, it is a summary in a high-dimensional space and not a description of the period's ideas. Third, since it is a centroid, it records the central tendency of the referential field rather than its full internal structure; the landscape, in other words, is only ever approached through its mean. Bearing this scope in mind, by computing semantic profiles for each century and then measuring the cosine distance between them, it becomes possible to ask whether centuries that are temporally close (e.g., 5th and 4th century BCE) tend to be close in semantic content too, or whether the conceptual worlds referenced by biographies shift dramatically across time, revealing discontinuities in the cultural and intellectual terrain that may not align with chronological proximity (see *1.6. Research Questions*).

Bibliography

1. Schich M, Song C, Ahn Y-Y, et al (2014) A network framework of cultural history. Science 345:558–562. https://doi.org/10.1126/science.1240064

2. Goldfarb D, Merkl D, Schich M (2015) Quantifying Cultural Histories via Person Networks in Wikipedia

3. Manovich L (2016) The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics. jca 1:1208. https://doi.org/10.22148/16.004

4. Suarez J-L, McArthur B, Soto-Corominas A (2015) Cultural Networks and the Future of Cultural Analytics. In: 2015 International Conference on Culture and Computing (Culture Computing). IEEE, Kyoto, Japan, pp 95–98

5. Ahnert R, Ahnert SE, Coleman CN, Weingart SB (2020) The Network Turn: Changing Perspectives in the Humanities, 1st ed. Cambridge University Press

6. Miccio LA, Agapitos P, Gamez-Perez C, et al (2025) Wikipedia as a cultural lens: a quantitative approach for exploring cultural networks. Humanit Soc Sci Commun 12:462. https://doi.org/10.1057/s41599-025-04772-5

7. Aragón P, David Laniado, Laniado D, et al (2012) Biographical social networks on Wikipedia: a cross-cultural study of links that made history. WikiSym '12 19. https://doi.org/10.1145/2462932.2462958

8. Ibrahim M, Danforth CM, Dodds PS (2017) Connecting every bit of knowledge: The structure of Wikipedia's First Link Network. Journal of Computational Science 19:21–30. https://doi.org/10.1016/j.jocs.2016.12.001

9. Gabella M (2019) Cultural Structures of Knowledge from Wikipedia Networks of First Links. IEEE Trans Netw Sci Eng 6:249–252. https://doi.org/10.1109/TNSE.2018.2812788

10. Jatowt A, Kawai D, Tanaka K (2016) Digital History Meets Wikipedia: Analyzing Historical Persons in Wikipedia. In: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries. Association for Computing Machinery, New York, NY, USA, pp 17–26

11. Ju H, Zhou D, Blevins AS, et al (2022) Historical growth of concept networks in Wikipedia. Collective Intelligence 1:263391372211098. https://doi.org/10.1177/26339137221109839

12. Rollin G, Lages J (2025) The most influential philosophers in Wikipedia: a multicultural analysis. EPJ Data Sci 14:69. https://doi.org/10.1140/epjds/s13688-025-00570-w

13. Schwartz GA (2021) Complex networks reveal emergent interdisciplinary knowledge in Wikipedia. Humanit Soc Sci Commun 8:1–6. https://doi.org/10.1057/s41599-021-00801-1

14. Miccio LA, Gámez-Pérez C, Suárez JL, Schwartz GA (2022) Mapping the networked context of copernicus, michelangelo, and della mirandola in wikipedia. Advs Complex Syst 25:2240010. https://doi.org/10.1142/S0219525922400100

15. Milne D, Witten IH (2008) An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In: Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy. AAAI Press, Chicago, USA

16. Cilibrasi RL, Vitanyi PMB (2007) The Google Similarity Distance. IEEE Trans Knowl Data Eng 19:370–383. https://doi.org/10.1109/TKDE.2007.48

17. Burns WE (2016) The scientific revolution in global perspective. Oxford University Press, New York

18. Scholten A, Smeets R (2024) Peripheral Networks: Canon-Formation in the Nineteenth-Century Reception of Regionalist Writers. Dutch Crossing 48:31–57. https://doi.org/10.1080/03096564.2023.2269789

19. Ruprechter T, Santos T, Helic D (2020) Relating Wikipedia article quality to edit behavior and link structure. Appl Netw Sci 5:61. https://doi.org/10.1007/s41109-020-00305-y

20. Schwartz GA (2021) Complex networks reveal emergent interdisciplinary knowledge in Wikipedia. Humanit Soc Sci Commun 8:127. https://doi.org/10.1057/s41599-021-00801-1

21. Miccio LA, Gámez-Pérez C, Suárez JL, Schwartz GA (2022) Mapping The Networked Context Of Copernicus, Michelangelo, and Della Mirandola in Wikipedia. Advs Complex Syst 25:2240010. https://doi.org/10.1142/S0219525922400100

22. Coscia M (2021) The Atlas For The Aspiring Network Scientist. Michele Coscia

23. Dimitrov D, Singer P, Lemmerich F, Strohmaier M (2017) What Makes a Link Successful on Wikipedia? In: Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, Perth Australia, pp 917–926

24. Lamprecht D, Lerman K, Helic D, Strohmaier M (2017) How the structure of Wikipedia articles influences user navigation. New Review of Hypermedia and Multimedia 23:29–50. https://doi.org/10.1080/13614568.2016.1179798

25. Bastian M, Heymann S, Jacomy M (2009) Gephi: An Open Source Software for Exploring and Manipulating Networks. ICWSM 3:361–362. https://doi.org/10.1609/icwsm.v3i1.13937

26. Fruchterman TMJ, Reingold EM (1991) Graph drawing by force-directed placement. Softw Pract Exp 21:1129–1164. https://doi.org/10.1002/spe.4380211102

27. Newman MEJ (2003) The Structure and Function of Complex Networks. SIAM Rev 45:167–256. https://doi.org/10.1137/S003614450342480

28. Gel YR, Lyubchich V, Ramirez Ramirez LL (2017) Bootstrap quantification of estimation uncertainties in network degree distributions. Sci Rep 7:5807. https://doi.org/10.1038/s41598-017-05885-x

29. Thompson ME, Ramirez Ramirez LL, Lyubchich V, Gel YR (2016) Using the bootstrap for statistical inference on random graphs. Can J Statistics 44:3–24. https://doi.org/10.1002/cjs.11271

30. Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman & Hall, New York

31. Jurafsky D, Martin JH (2026) 4.11.1 The Paired Bootstrap Test. In: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd ed. pp 88–89

32. Agapitos P, Suárez J-L, Schwartz GA (2025) WikiTextGraph: A Python Tool for Parsing Multilingual Wikipedia Text and Graph Extraction. Journal of Open Research Software 13:17. https://doi.org/10.5334/jors.572

33. Attardi G (2025) attardi/wikiextractor

34. Wenger JC (2024) jchwenger/dataset.wikimedia-cirrus

35. Studer W (2023) studerw/wiki-dump-parser

36. (2025) wikimedia/pywikibot

37. Consonni C, Laniado D, Montresor A (2019) WikiLinkGraphs: A complete, longitudinal and multi-language dataset of the Wikipedia link networks. CoRR abs/1902.04298:

38. Aspert N, Miz V, Ricaud B, Vandergheynst P (2019) A Graph-Structured Dataset for Wikipedia Research. In: Companion Proceedings of The 2019 World Wide Web Conference. ACM, San Francisco USA, pp 1188–1193

39. Reznik I, Shatalov V (2016) Hidden revolution of human priorities: An analysis of biographical data from Wikipedia. Journal of Informetrics 10:124–131. https://doi.org/10.1016/j.joi.2015.12.002

40. Viseur R (2013) Extraction of Biographical Data from Wikipedia. In: Proceedings of the 2nd International Conference on Data Technologies and Applications. SciTePress - Science and and Technology Publications, Reykjavík, Iceland, pp 248–252

41. Moreira J, Neto EC, Barbosa L (2021) Analysis of structured data on Wikipedia. IJMSO 15:71. https://doi.org/10.1504/IJMSO.2021.117108

42. Heist N, Paulheim H (2020) Entity Extraction from Wikipedia List Pages. In: Harth A, Kirrane S, Ngonga Ngomo A-C, et al (eds) The Semantic Web. Springer International Publishing, Cham, pp 327–342

43. Eom Y-H, Aragón P, Laniado D, et al (2015) Interactions of Cultures and Top People of Wikipedia from Ranking of 24 Language Editions. PLOS ONE 10:e0114825. https://doi.org/10.1371/journal.pone.0114825

44. Wikidata. https://www.wikidata.org/?uselang=en. Accessed 21 Apr 2026

45. Humaniki | Wikimedia Diversity Dashboard Tool. https://humaniki.wmcloud.org/. Accessed 2 Mar 2026

46. Huq KT, Ciampaglia GL (2024) Survival of the Notable: Gender Asymmetry in Wikipedia Collective Deliberations

47. Qi P, Zhang Y, Zhang Y, et al (2020) Stanza: A Python Natural Language Processing Toolkit for Many Human Languages

48. Reznik I, Shatalov V (2016) Hidden revolution of human priorities: An analysis of biographical data from Wikipedia. Journal of Informetrics 10:124–131. https://doi.org/10.1016/j.joi.2015.12.002

49. Lerner J, Lomi A (2018) Knowledge categorization affects popularity and quality of Wikipedia articles. PLoS ONE 13:e0190674. https://doi.org/10.1371/journal.pone.0190674

50. Reimers N, Gurevych I (2019) Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks

51. Jurafsky D, Martin JH (2024) Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition

52. Deb S, Chanda AK (2022) Comparative analysis of contextual and context-free embeddings in disaster prediction from Twitter data. Machine Learning with Applications 7:100253. https://doi.org/10.1016/j.mlwa.2022.100253

53. Liu Y, Tilahun G, Gao X, et al (2024) Comparative Analysis of Static and Contextual Embeddings for Analyzing Semantic Changes in Medieval Latin Charters

54. Ethayarajh K (2019) How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp 55–65

55. Eom Y-H, Shepelyansky DL (2013) Highlighting entanglement of cultures via ranking of multilingual Wikipedia articles. PLOS ONE 8:1–10. https://doi.org/10.1371/journal.pone.0074554

56. Cristian Jara-Figueroa, Jara-Figueroa C, Jara-Figueroa C, et al (2015) How the medium shapes the message: Printing and the rise of the arts and sciences. arXiv: Computers and Society. https://doi.org/10.1371/journal.pone.0205771

57. Kozlowski AC, Taddy M, Evans JA (2019) The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. Am Sociol Rev 84:905–949. https://doi.org/10.1177/0003122419877135

58. Stoltz DS, Taylor MA (2021) Cultural cartography with word embeddings. Poetics 88:101567. https://doi.org/10.1016/j.poetic.2021.101567

59. Kirchner C, Mohr JW (2010) Meanings and relations: An introduction to the study of language, discourse and networks. Poetics 38:555–566. https://doi.org/10.1016/j.poetic.2010.09.006

60. Mohr JW (1998) Measuring Meaning Structures. Annu Rev Sociol 24:345–370. https://doi.org/10.1146/annurev.soc.24.1.345

61. Chakrabarti P, Frye M (2017) A mixed-methods framework for analyzing text data: Integrating computational techniques with qualitative methods in demography. DemRes 37:1351–1382. https://doi.org/10.4054/DemRes.2017.37.42

62. Stoltz DS (2024) Mapping Texts: Computational Text Analysis for the Social Sciences. Oxford University Press, Incorporated, Oxford

63. Lee M, Martin JL (2015) Coding, counting and cultural cartography. Am J Cult Sociol 3:1–33. https://doi.org/10.1057/ajcs.2014.13

64. Jurafsky D, Martin JH (2026) Embeddings. In: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 3rd ed. pp 96–119

65. Alkaabi H, Jasim AK, Darroudi A (2025) From Static to Contextual: A Survey of Embedding Advances in NLP. PERFECT 2:57–66. https://doi.org/10.62671/perfect.v2i2.77

3.1. Wikipedia as a Cultural Lens – a Network Approach

3.1.1. Background

The conceptual (see *Introduction*) and methodological (see *Methodology - section 2.2*) foundations established in the previous chapters place this thesis within digital humanities and cultural analytics and identify Wikipedia's wikilinks network as the empirical object of analysis. The present section presents the results.[23] But first, it is worth situating the specific scholarly conversation this thesis contributes to.

The first body of literature relevant to the present section concerns the use of large-scale datasets to investigate questions traditionally framed within the humanities and social sciences. This line of work generally recognizes that cultural dynamics (e.g., the formation of movements, the diffusion of ideas, the rise and decline of intellectual currents) emerges from interactions among many people, ideas, and cultural objects, and that those interactions are increasingly accessible in digital form. Schich et al. [2] mapped the migration trajectories of notable individuals across two millennia and using the aggregate movement of people they characterized the geographic dynamics of cultural transmission. Another study adopts a different strategy by analysing the frequencies with which specific terms appear in millions of digitised books to track lexical and conceptual change at unprecedented scale [3]. Other studies extend this orientation in different directions: for instance, Gao et al. [4] apply tools from statistical physics to identify long-range correlations in cultural and social phenomena over the past two centuries, while another paper argues for a more general programme, sometimes referred to as *cliodynamics*, that treats historical processes as analytically

[23] The results presented in this section has been published as an independent peer-reviewed paper with the title "Wikipedia as a cultural lens: a quantitative approach for exploring cultural networks" in the journal of *Humanities & Social Sciences Communications* in 2025 [1].

tractable through formal modelling [5]. Even though these approaches are different from each other, the unifying component can be tracked in the underlying premise that cultural patterns, once aggregated across large populations of cultural entities, exhibit regularities that remain invisible at the level of any single case. This is exactly the premise that motivates the present analysis.

A second body of literature, complementary to the first, formalizes the analysis of cultural systems through the tools of complex network theory. A central claim of this literature is that the structure of relationships among cultural entities is itself a source of information about how those entities behave collectively, what are the patterns that emerge when relationships are abstracted, and how the system affects the connections between the entities. For instance, Smolla and Akçay [6] use complex network analysis to study cultural evolution, demonstrating how individual behaviour at the micro level gives rise to structural patterns observable at the population level. By applying a related logic to the early modern publication industry, another study [7] reconstructed the social network underlying book production from large-scale archival data. Focusing mostly on the theoretical framework of these types of works, Suárez et al [8] introduced the concept of a cultural network, that is defined as a structure in which relationships among people are mediated by cultural entities such as ideas, artefacts, and institutions rather than by direct social ties alone. Such networks constitute a distinct empirical object that warrants its own analytical framework. This orientation is generalised under the heading of the network turn in the humanities (see *Introduction*), arguing that network analysis offers a shared conceptual vocabulary capable of bringing computational rigour into dialogue alongside humanistic interpretation [9].

The significance of the aforementioned works lies in two specific contributions. First, the recognition that cultural networks are not merely social networks with cultural content; they are structurally distinct because they can also carry connections between cultural entities and not only interpersonal ones. In other words, these studies shift the analytical focus from who connects to whom toward which cultural entities (e.g., ideas,

people, works, concepts) bridge different domains of knowledge. The second contribution is that the toolkit that network science offers provide a quantitative vocabulary for describing structural properties that humanistic methods have discussed before in qualitative terms. For example, notions such as influence, isolation, bridging acquire formal counterparts that can be measured, compared, and tested.

The third strand of relevant literature concerns the use of Wikipedia as an empirical resource for quantitative cultural research. As mentioned in the *Introduction*, Wikipedia is the main data source of this thesis and it satisfies three conditions that make it analytically distinctive: 1) it operates at a scale that requires computational methods to be analysed, 2) it is structured given the guidelines the editors have to follow, and 3) it is produced through sustained collective editorial negotiation. As is normal, a significant body of work has exploited these properties for a variety of analytical purposes. For example, Mehdi et al. [10] provide a systematic review of studies that use the Wikipedia corpus, identifying applications in information retrieval, natural language processing, and ontology construction. More recently, it has been used to answer cultural questions. For instance, the Wikipedia corpus has been used to examine its role in representing cultural heritage [11], as an issue in the *Journal of Cultural Analytics* to map the structure of world literature [12], to document gender and country biases embedded in its citation patterns [13]. A subset of studies shifting the focus from citation patterns to the wikilinks of Wikipedia themselves, treats the structure of internal connections as the primary analytical object. These studies are the ones that are most directly relevant to the present chapter. In the studies below, Wikipedia's network is used to investigate specific historical periods through targeted, period-focused cultural networks. For instance Eom and Shepelyansky [14] and Eom et al. [15] use the ranking of articles across multiple language editions to identify cultural entanglement among national perspectives. A related approach constructed co-citation networks among scientific topics represented in Wikipedia to map the open knowledge structure of Science [16]. Finally, two related studies examine the cultural

structures encoded in the network of first-wikilinks [17, 18]. More recent works, develop methodologies proposing that Wikipedia's network can be converted into a meaningful network of knowledge by means of a structural relationship metric [19]. In that study, the cultural contexts of Pablo Picasso, Albert Einstein, and James Joyce are analysed jointly to characterise the interdisciplinary structure of early 20th-century culture across Art, Science, and Literature. Similarly, a more recent work, focusing on the Italian Renaissance, took Michelangelo, Copernicus, and Pico della Mirandola as a representative triad of artist, scientist and philosopher and identified hidden structural relationships between the nodes of this network [20].

The literature reviewed above establishes three points that motivate chapter's analysis. First, quantitative analysis of cultural data has been shown to recover patterns that traditional humanistic methods, working usually at a small scale, cannot easily detect. Second, complex network theory provides a formal apparatus for analysing such patterns when they take the form of structured relationships among cultural entities. Third, Wikipedia's network, when interpreted as a cultural network, constitutes a particularly well-suited empirical source for this kind of analysis. What remains absent from the literature, however, is an analysis that moves toward the statistical characterisation of a historical period and distances itself as much as possible from the analysis of a few individuals.

The work presented in this chapter is designed to fill that gap. Instead of analysis a single triad of figures, it proposes to characterise the cultural network of a historical period by averaging across many triads sampled from that period (see *Methodology - section 2.2*). By doing so, it aims to recover the structural properties of the period itself, as distinct from those of any particular configuration of individuals within it. As a proof of concept, the cultural network of Art, Science, and Philosophy in 17th century Europe is used. The results presented in the following sections are evaluated against a dual criterion: 1) agreement with established historical knowledge, which means that the approach can recover cultural meaningful patterns, and 2) the disclosure of new

quantitative regularities that would not have been accessible through traditional non-computational methods alone.

3.1.2. Results

3.1.2.1. The Network as a Whole: Modularity of the Period

As mentioned earlier (see *Methodology - Modularity*), modularity measures how cleanly the cultural network divides into separate disciplinary clusters. It is high when nodes within a discipline connect densely to one another and sparsely to nodes in other disciplines. It is low when frequent cross-disciplinary links blur the boundaries between them. The value is normalised, so a score close to 1 indicates strong clustering and a score close to 0 weak clustering.

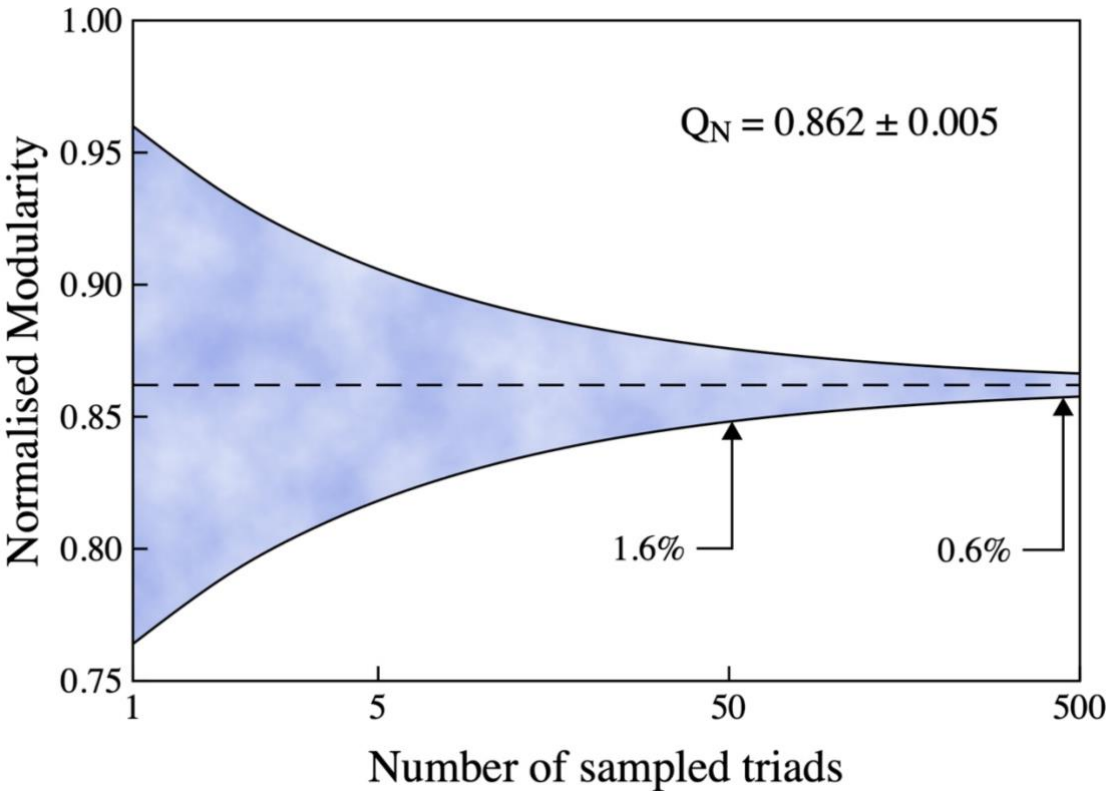


Figure 18 Average normalised modularity ($\overline{Q_N}$) and its associated error interval as a function of the number of sampled triads. The relative error decreases as the number of sampled triads increases, reaching 1.6% for 50 samples and 0.6% for 465 samples (the full dataset). The dashed line marks the mean value obtained using the bootstrapping method, $\overline{Q_N} = 0.862 \pm 0.005$. Note the logarithmic scale on the horizontal axis.

Figure 18 reports the average normalised modularity for the 17th-century network as a function of the number of sampled triads. Two observations emerge from this plot. First, the estimated error decreases substantially with the number of triads. As it can be seen in Figure 18, the error stands at approximately 1.6% for 50 triads and falls at 0.6% when all 465 triads are included. This is a finding that confirms that the chosen sample size is sufficient for reliable estimation. Second, the average value of the modularity stabilises at $\overline{Q_N} = 0.862 \pm 0.005$, with a standard deviation of 0.098, which is approximately 11.4% of the mean.[24] This indicates that the 17th century network divides the cluster cleanly into three disciplinary clusters, and that high modularity is a persistent feature across the sampled triads rather than being driven by a few exceptional cases; though the standard deviation of 0.098 (i.e., 11.4% of the mean) does reflect a moderate degree of variability in the strength of that separation from triad to triad.

The interpretation of this result can be aided when placed next to prior studies that applied similar methodology to other historical periods. For the Italian Renaissance, for example, the cultural network built around Michelangelo, Copernicus, and Pico della Mirandola produced a normalised modularity of 0.77 [20]. For the early 20th-century network built around Picasso, Einstein, and Joyce, the corresponding value was 0.88 [19]. Even though these earlier studies analyse single triads and not period averages, the progression from 0.77 (Italian Renaissance) to 0.862 (17th century) to 0.88 (early 20th century) is consistent with the well-established historical hypothesis that disciplinary specialization has increased progressively across the modern period [21]. The disciplines, in other words, have become internally cohesive and more externally separated over time. The small number of comparison points between the aforementioned period can only confirm a suggestive trend, nevertheless it provides

[24] The overbar in $\overline{Q_N}$ is used here to indicate that the value is averaged across many triads and therefore characterises the period as a whole, in contrast to the single-network modularities reported in earlier studies of specific triads [19, 20](which are written without the overbar as Q_N and refer to one particular configuration of individuals).

initial evidence that the method is sensitive to historically meaningful structural change.

To sum up, this first finding supports the claim that was made at the beginning of this chapter: that Wikipedia's network encodes culturally meaningful structural information that can be recovered using complex network theory in combination with the NGD (see *Methodology - Structural Relationships using the Normalised Google Distance*).

3.1.2.2. Disciplines and their Interactions

Modularity establishes that the network divides into clusters, but it does not specify neither how strong the internal cohesion of each cluster is, nor how the clusters interact with one another. For this reason, assortativity is employed (see *Methodology - Assortativity*). The assortativity matrix provides this finer-grained information. Its diagonal elements (a_{ii}^{N}) measure the strength of connections within each cluster (i.e., how tightly Art connects to Art, Science to Science, and Philosophy to Philosophy). Its off-diagonal elements (a_{ij}^{N}, where $i \neq j$) measure the strength of connections between clusters (i.e., how strongly Art connects to Science, Art to Philosophy, and Science to Philosophy). The matrix is normalised so that each coefficient represents the fraction of weighted links accounted for by that particular intra- or inter-cluster relationship.

Table 1 Normalised assortativity coefficients ($\bar{\alpha}_{ij}^{N}$) and their bootstrapping error estimates for each pair of disciplines. The standard deviation (σ_{ij}) of each bootstrapped distribution is also reported, with the percentage in parentheses indicating its magnitude relative to the corresponding mean. Diagonal entries represent within-discipline assortativity, while off-diagonal entries capture cross-discipline mixing. Row and column colours correspond to Art (red), Science (green), and Philosophy (blue).

Assortativity coefficients	*Art*	*Science*	*Philosophy*
Art	$\overline{\alpha_{11}^{N}} = 0.285 \pm 0.004$ $\sigma_{11} = 0.085\ (30\%)$	$\overline{\alpha_{12}^{N}} = 0.025 \pm 0.002$ $\sigma_{12} = 0.035\ (140\%)$	$\overline{\alpha_{13}^{N}} = 0.017 \pm 0.001$ $\sigma_{13} = 0.020\ (117\%)$
Science		$\overline{\alpha_{22}^{N}} = 0.268 \pm 0.003$ $\sigma_{22} = 0.074\ (28\%)$	$\overline{\alpha_{23}^{N}} = 0.092 \pm 0.003$ $\sigma_{23} = 0.065\ (71\%)$
Philosophy			$\overline{\alpha_{33}^{N}} = 0.313 \pm 0.005$ $\sigma_{33} = 0.115\ (37\%)$

Intra-disciplinary interactions

Starting with the intra-cluster connectivity, the diagonal coefficients reveal that the three disciplines are not equally cohesive internally. For 465 triads, the values are $\overline{\alpha_{11}^{N}} = 0.285 \pm 0.004$ for Art, $\overline{\alpha_{22}^{N}} = 0.268 \pm 0.003$ for Science, and $\overline{\alpha_{33}^{N}} = 0.313 \pm 0.005$ for Philosophy (see **Table 1** and **Figure 19 (left)** for the corresponding error intervals when increasing the number of triads). It is worth noting that the differences between the three coefficients are larger than their respective combined standard errors which strongly suggests that the disciplinary differences observed are not artefacts of sampling but reflect a structural variation in intra-cluster cohesion. Moreover, Philosophy emerges as the discipline with the most internal-to-philosophy links, with

an intra-cluster coefficient roughly 17% higher than that of Science and slightly higher than that of Art.

This finding is historically distinctive when compared with prior results. In both the Italian Renaissance ($\alpha_{11}^{N} = 0.383$) and the early 20th century ($\alpha_{11}^{N} = 0.316$), Art was the most internally cohesive cluster [19, 20]. The 17th-century network thus marks a departure from this pattern, with Philosophy taking the position previously occupied by Art. This points to a structural reconfiguration of the relations among the three disciplines in the seventeenth century, which the inter-cluster coefficients help to specify further.

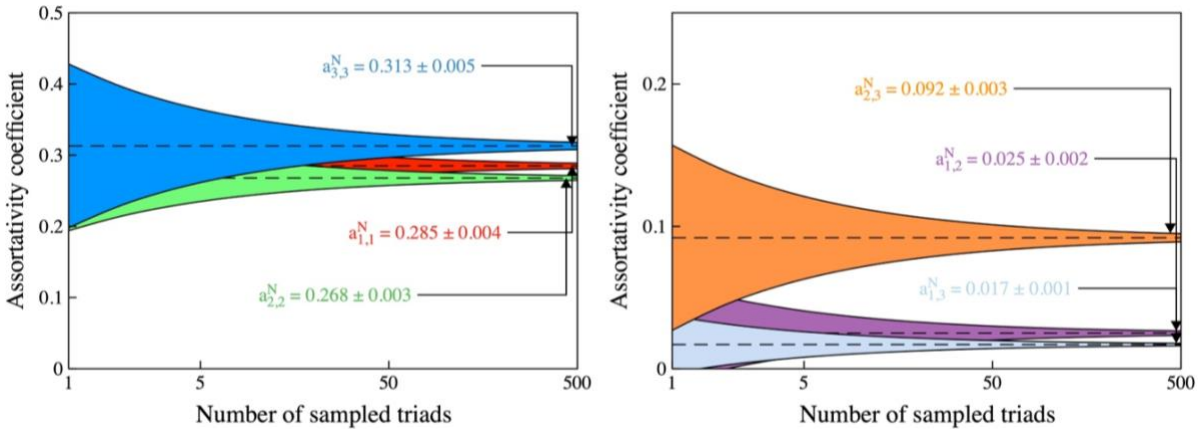


Figure 19 Average normalised assortativity coefficients as a function of the number of sampled triads, estimated using the bootstrapping method. Shaded regions represent the associated error intervals, which narrow as the number of sampled triads increases. **(Left)** Diagonal coefficients ($\overline{\alpha_{ii}^{N}}$), reflecting the strength of within-discipline interactions for Art (red), Science (green), and Philosophy (blue). **(Right)** Off-diagonal coefficients ($\overline{\alpha_{jj}^{N}}$, $i \neq j$), reflecting the strength of cross-discipline interactions between Art and Science (violet), Art and Philosophy (light blue), and Science and Philosophy (orange). The dashed lines and cross markers indicate the mean values obtained using the full dataset. Note the logarithmic scale on the horizontal axis.

Inter-disciplinary interactions

While I focused on diagonal coefficients to measure the strength within the disciplines of the network, now I will focus on the off-diagonal coefficients which quantify the strength of connections between different disciplines. The three relevant pairs return very different values: $\overline{\alpha_{12}^{N}} = 0.025 \pm 0.002$ for Art–Science, $\overline{\alpha_{13}^{N}} = 0.017 \pm 0.001$ for Art–Philosophy, and $\overline{\alpha_{23}^{N}} = 0.092 \pm 0.003$ for Science–Philosophy. All three values are

substantially smaller than the intra-cluster coefficients, which is expected given the high modularity of the network. However, the difference in the relative magnitudes among them is an interesting finding.

The Science-Philosophy connection ($\overline{\alpha_{23}^{N}} = 0.092 \pm 0.003$) is the strongest of the three, exceeding the Art-Science connection by a factor of roughly four and the Art-Philosophy connection by a factor of more than five. This pattern aligns with the historical literature on the period. Sandoz [22] has documented that early modern Philosophy underwent a substantial reorientation during the 17th century, shifting its attention away from questions of Art and aesthetics and towards epistemology, the methodology of science, and the foundations of scientific knowledge. Thomas Hobbes captures this shift in *Leviathan* (1651) [23], where he distinguishes between "Knowledge of Fact" (the kind required of a witness) and "Knowledge of the Consequence" of one statement to another (the kind required of a philosopher), explicitly aligning the latter with Science [22]. The strong Science-Philosophy coefficient (see **Figure 19** (right panel-orange)) recovered here is consistent with this re-orientation: the cultural network reflects a period in which the two disciplines were overlapping and only gradually separating around questions of mind and matter, the foundations of the scientific method, and the nature of the physical explanation [24, 25].

The Art-Science and Art-Philosophy coefficients are also worth comparing with prior results. The Art-Science coefficient in the 17th century ($\overline{\alpha_{12}^{N}} = 0.025$) (see **Figure 19** (right panel-violet)) is similar to that observed at the beginning of the 20th century ($a_{12}^{N} = 0.020$) [19] but substantially lower than the value found for the Renaissance ($a_{12}^{N} = 0.056$) [20]. This difference aligns with the established historical claim that the Art-Science relationship was unusually close during the Italian Renaissance and weakened in the centuries that followed [26]. The Art-Philosophy coefficient shows an even more pronounced difference: the 17th century value ($\overline{\alpha_{13}^{N}} = 0.017$) (see **Figure 19** (right panel-light blue)) is approximately one-eighth of the Italian Renaissance value

($a_{13}^{N} = 0.130$). If we take together these comparisons, we can see that the Art cluster has progressively decoupled from both Science and Philosophy since the Italian Renaissance, while the Science-Philosophy connections has strengthened. Thus, the cultural network of Wikipedia captures a structural transformation in how the three disciplines examined here relate to one another, and that transformation is consistent with what historical scholarship has independently discussed.

From an analytical point of view, another interesting property concerns the variability of these coefficients across triads, measured by the relative standard deviation (i.e., the standard deviation divided by the mean, expressed as a percentage). For the diagonal coefficients, this relative variability sits at approximately 30% (see **Table 1**), indicating relatively consistent intra-cluster behaviour across triads. For the off-diagonal coefficients, however, the values are higher: 71% for Science–Philosophy, 117% for Art–Philosophy, and 140% for Art–Science. This means that the Science–Philosophy connection is not only the strongest but also the most homogeneous, while the connections involving Art vary substantially depending on which artist is chosen as a seed. This heterogeneity is itself historically meaningful: the 17th century Art world was structured by regional fragmentation, with separate channels of production and patronage operating for Protestant and Catholic markets in Europe and Latin America [27] and with local and national patrons shaping the development of distinct national art traditions [28]. The cultural network registers this fragmentation as elevated variance in the Art-related inter-cluster coefficients.

Distributions of the Coefficients

Beyond their mean values, the coefficients also follow distinct statistical distributions that reveal something more about how the network is organised. **Figure 20** (left panel) shows the distribution of $\overline{\alpha_{11}^{N}}$ values across the 465 triads, which is well described by a lognormal function. The other diagonal coefficients follow similar distributions. By contrast, Figure 20 (right panel) shows that the distribution of $\overline{\alpha_{12}^{N}}$ values follows a

power law with an exponent of approximately -2 (see **Figure 20** (right panel)). A power-law distribution differs from a lognormal one in that it features a long, heavy tail: most triads produce low values, but a small number of triads produce values much higher than the average. This distinction matters because it indicates that intra-cluster and inter-cluster connections are generated by different underlying processes, a point that becomes more concrete in the analysis of individual nodes below.

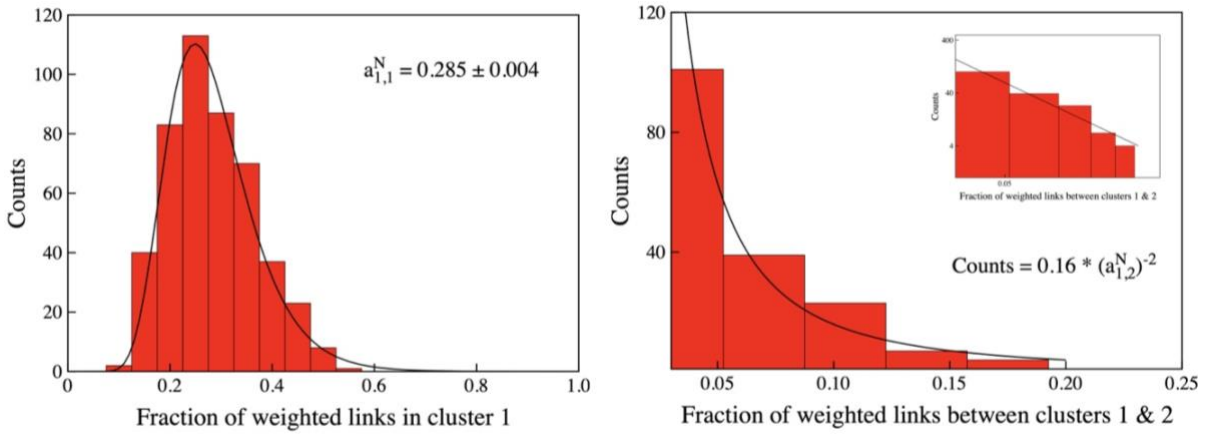


Figure 20 Bootstrapped distributions of the normalised assortativity coefficients for the within-discipline term $\overline{\alpha_{11}^N}$ (**left**) and the cross-discipline term $\overline{\alpha_{12}^N}$ (**right**). The solid curves show the best fits to each distribution: a log-normal function for $\overline{\alpha_{11}^N}$ and a power law of the form Counts $= 0.16 \times (\overline{\alpha_{12}^N})^{-2}$ for $\overline{\alpha_{12}^N}$. The inset in the right panel shows the same data on a double-logarithmic scale, where the linear trend confirms the power-law behaviour.

Openness of the Disciplines

Continuing with the discipline-level metrics, a complementary perspective on that aspect is provided by openness, which is defined as the percentage of nodes within a discipline whose external connections (i.e., to other disciplines) outweigh their internal connections (i.e., within the same discipline). A cluster with high openness contains many nodes that act as bridges to other disciplines; a cluster with low openness is more inwardly turned (see *Methodology - Openness*).

Our analysis shows that for the 17th century cultural network, the average openness are $\overline{Op_{art}} = 5.0 \pm 0.4\%$, $\overline{Op_{scl}} = 15.5 \pm 1.0\%$, and $\overline{Op_{phil}} = 13.4 \pm 1.1\%$. The results show that Art is the most inwardly oriented of the three disciplines, with an openness

approximately three times lower that than of Science or Philosophy. This finding is consistent with the inter-cluster assortativity coefficients reported above (see **Table 2**), which showed that Art connects only weakly to both Science and Philosophy.

Table 2 Average openness ($\overline{Op}$) for each discipline, estimated using the bootstrapping method. The standard deviation (σ) of each bootstrapped distribution is given below the mean, with the percentage in parentheses indicating its magnitude relative to the corresponding mean. The large relative standard deviations (143–178%) reflect the high variability inherent to the openness measure, which is sensitive to the choice of seed node in each triad.

Art	*Science*	*Philosophy*
$\overline{Op} = 5.0 \pm 0.4\%$ $\sigma = 8.1\ (162\%)$	$\overline{Op} = 15.5 \pm 1.0\%$ $\sigma = 22.1\ (143\%)$	$\overline{Op} = 13.4 \pm 1.1\%$ $\sigma = 23.8\ (178\%)$

Moreover, when examining the distribution of openness values (see **Figure 21**), it is shown that it follows a power law, like that of the inter-cluster coefficients. For the Art cluster, the exponent is $\alpha = -1.14$; for the Science cluster, $\alpha = -0.95$; and for the Philosophy cluster, $\alpha = -1.09$. The interpretation of these values tells us that overall, the majority of nodes within any given discipline are weakly connected to nodes in other clusters, while a small number of nodes are very strongly connected across discipline boundaries. This shows a signal of homophily, which is the tendency of nodes to associate preferentially with similar nodes, but it also points to the existence of small number of strongly cross-connecting nodes that function as bridges between disciplines. These bridging nodes are examined more closely in the analysis at the individual-node level that follows.

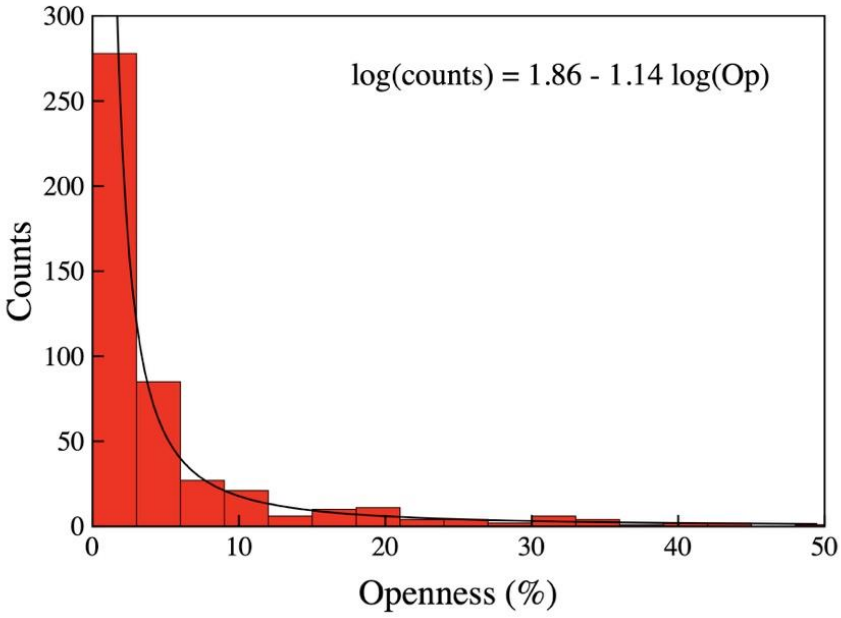


Figure 21 Bootstrapped distribution of openness values for the Art discipline. The distribution is strongly right skewed, with the majority of triads exhibiting low openness values. The solid curve shows the best power-law fit to the data, given by $\log(\text{Counts}) = 1.86 - 1.14\log(Op)$, confirming that large openness values are rare but present.

This structure invites comparison with Granovetter's "strength of weak ties" thesis. Granovetter observed that cross-group[25] connections in social networks tend to be weak ties, and that these weak ties are disproportionately responsible for transmitting novel information across otherwise homophilous communities [29]. The distribution observed here is consistent with the underlying geometry of that thesis since I observe dense intra-discipline association alongside sparse inter-discipline connectivity. However, the findings presented here must be connected to Granovetter's thesis with some caution. The power-law tail indicates that inter-discipline connectivity is not uniformly distributed across weakly bridging nodes, but it appears concentrated in a comparatively small portion of the network where connectivity across discipline boundaries is robust rather than weak. With this in mind, I argue that these nodes correspond more closely to what Burt and Soda characterized as brokers spanning structural holes; in other words, positions whose value derives precisely from

[25] By "group" here we can think of disciplines, such as the ones presented in this section (Art, Philosophy, Science).

maintaining robust ties to otherwise disconnected communities [30]. The 17th century cultural network thus exhibits a hybrid bridging regime, in which the bulk of inter-discipline traffic is carried, in aggregate, by many sparse cross-ties, while a share of integrative links appears concentrated in a small portion of the network that functions as bridges between disciplines.

Table 3 Openness values for representative triads, grouped by which seed node was varied. In the first eight entries, the Science and Philosophy seeds are held fixed while the Art seed is varied; the openness values associated with Newton and Spinoza remain relatively stable across different Art seeds, suggesting that the choice of Art representative has little influence on these disciplines. In the final four entries, the Art and Philosophy (Phil.) seeds are held fixed while the Science (Sci.) seed is varied; replacing Huygens with Newton produces little change in openness, whereas substituting Hevelius leads to a substantial change, illustrating the sensitivity of the measure to the choice of Science seed.

Triads' seeds			**Openness (Op)**		
Art	*Sci.*	*Phil.*	*Art*	*Sci.*	*Phil.*
Nicolaes Maes	Isaac Newton	Baruch Spinoza	1.79	0	88.98
Claude Lorrain	Isaac Newton	Baruch Spinoza	3.81	0	88.70
Johannes Vermeer	Isaac Newton	Baruch Spinoza	3.05	0	90.91
Peter Lely	Isaac Newton	Baruch Spinoza	26.67	0.23	84.44
Christopher Wren	Isaac Newton	Baruch Spinoza	31.96	0.49	77.33
Gerrit Dou	Isaac Newton	Baruch Spinoza	3.75	0	90.91
Andrea Pozzo	Isaac Newton	Baruch Spinoza	4.08	0	87.50
Godfrey Kneller	Isaac Newton	Baruch Spinoza	37.25	0.42	88.71

Johannes Vermeer	Christian Huygens	Blaise Pascal	3.70	0	43.97
Johannes Vermeer	Isaac Newton	Blaise Pascal	2.50	0	54.55
Johannes Vermeer	Johannes Hevelius	Blaise Pascal	0.63	10.61	0
Johannes Vermeer	Jan Swammerdam	Blaise Pascal	1.22	26.19	0

Openness also reveals another structural feature when examined at the level of specific triads. Table 3 shows openness values for triads in which two seeds are held constant, and one is varied. When Isaac Newton (Science) and Baruch Spinoza (Philosophy) are held fixed and different artists are substituted, the openness of the Newton-related clusters remains close to zero and the openness of the Spinoza-related cluster remains close to 90%, almost independently of the choice of the artist. This means that Newton's network position is largely insulated from Art connections, while Spinoza's position is strongly connected to the broader network (and, given the consistency of the pattern, primarily to the Newton-related cluster itself). This finding is consistent with what historians of science have established about the diffusion of Newtonian thought in continental Europe: the technical character of the *Principia* (1687) delayed its uptake outside England, and continental thinkers committed to Cartesianism resisted Newtonianism well into the 1740s [31, 32]. The cultural network registers this delay as elevated openness for the few artists with English connections (e.g., Peter Lely, who settled in London; Godfrey Kneller, who settled at that English court; and Christopher Wren, an English architect and natural philosopher) and very low openness for continental artists from Holland, France, and Italy.

However, a different pattern emerges when the Science seed is varied (see last four rows of **Table 3**). Replacing Newton with Huygens produces only modest changes in

discipline openness, but replacing Newton with Hevelius or Swammerdam produces larger changes, reflecting the fact that the latter two figures are positioned differently within the network of the 17th-century Science. These variations confirm that the method is sensitive to fine-grained differences in network position, even within a single discipline. This also means that openness values for any given triad are shaped by the specific seed chosen, and should be read as local, seed-dependent measurements rather than fixed properties of the discipline as a whole.

3.1.2.3. Individual Nodes: Frequencies and Connectivity Distributions

As mentioned in the *Introduction*, this thesis follows the analytical approach of focus reading. For this reason, an analysis at the level of individual nodes is necessary. This analysis serves two purposes: 1) it identifies which cultural entities are most central to the 17th century cultural network of Art-Science-Philosophy. 2) It characterises the statistical laws that govern how individual nodes connect to others, both within and across disciplines.

Before presenting the results of this part of the analysis, it should be noted that the 465 triads in this study produce networks with a total of 3,585 unique nodes. These nodes, as the representation of cultural entities, can be people (e.g., Blaise Pascal, Christiaan Huygens, Thomas Hobbes, etc.), objects (e.g., (microscope, telescope, vacuum pump, etc.), and ideas or concepts (e.g., rationalism, mechanical philosophy, scientific revolution, etc.).

Node Analysis: The Most Central Cultural Entities

The simplest analysis at the node level is to count how frequently each cultural entity appears across the 465 triads. Table 4 lists the entities that appear in more than half of the triads. Three observations can be made from this list.

First, most of the high-frequency nodes are not themselves seed entities. This means that they emerge as central not because they were selected at the outset, but because they are structurally important within the network of relations that the cultural entities collectively generate. The method therefore is able to capture relevant cultural entities that would not have been identified through a direct examination of the cultural entities list.

Table 4 Cultural entities appearing in more than half of the 465 sampled triads ($N \geq 233$), ranked by frequency N. Each entry corresponds to an English Wikipedia article and may represent a person, concept, or movement. The "discipline" column indicates whether the entity is primarily associated with Art, Science (Sci), Philosophy (Phil), or a combination thereof; entries without a birth–death date or discipline label correspond to concepts or movements rather than individuals. The dominance of Science and Philosophy among the most frequent nodes reflects the central role of these disciplines in the 17th-century cultural network, and notably, the majority of these entities were not used as triad seeds, demonstrating the ability of the method to recover implicit knowledge from the network.

Node's name	**Birth-Death**	**Discipline**	***N***
Pierre Gassendi	1592-1655	Sci-Phil	355
Robert Boyle	1627-1691	Sci	351
Mechanical Philosophy	-	-	338
René Descartes	1596-1650	Sci-Phil	326
Pierre Bayle	1647-1706	Phil	313
Nicolas Malebranche	1638-1715	Phil	313
Evangelista Torricelli	1608-1647	Sci	309
Rationalism	-	-	307
Giambattista Vico	1668-1744	Phil	299

Thomas Hobbes	1588-1679	Phil	296
Michel de Montaigne	1533-1592	Phil	296
Scientific revolution	-	-	294
Blaise Pascal	1623-1662	Sci-Phil	290
Marin Mersenne	1588-1648	Sci	276
Thomism	-	-	275
Baruch Spinoza	1632-1677	Phil	274
Gottfried Wilhelm Leibniz	1646-1716	Sci-Phil	263
Antoine Arnauld	1612-1694	Phil	263
Jean le Rond d'Alembert	1717-1783	Sci-Phil	262
Christiaan Huygens	1629-1695	Sci	261
Frans van Schooten	1615-1660	Sci	256
Gerolamo Cardano	1501-1576	Sci-Phil	255
Francis Bacon	1561-1626	Phil	252
Johannes Kepler	1571-1630	Sci	251
David Teniers the Younger	1610-1690	Art	249

Willem van de Velde the Younger	1633-1707	Art	241
Juan Caramuel y Lobkowitz	1606-1682	Phil	234

Second, the disciplines of Science and Philosophy seem to dominate the list of central cultural entities. Among the named individuals, philosophers, and people that worked between Science and Philosophy in the 17th century (e.g., René Descartes, Gottfried Wilhelm Leibniz, Blaise Pascal, etc.) substantially outnumber artists. Among the abstract concepts and ideas, the most prominent are "Mechanical philosophy", "Rationalism", "Scientific revolution", and "Thomism" (all of which sit at the intersection of Science and Philosophy). This finding reinforces the structural pattern identified through the assortativity matrix: within the tripartite network of Science, Philosophy, and Art that we extract from English Wikipedia, the 17th century is organised around the axis of the first two, while Art occupies a more peripheral position relative to them.

Third, the most central nodes correspond to people and ideas rather than to physical objects. Microscope, telescope, and vacuum pump, albeit scientific instruments of the 17th century Science, do not appear among the high-frequency nodes. This suggests that, in Wikipedia's representation of the period, the relational structure is organised around conceptual and biographical entities rather than around material culture. Whether this reflects a property of the period itself or a property of how Wikipedia documents it is a question to which the analysis returns in the discussion of limitations (see *Limitations*).

Intra-Cluster Connectivity: A Core-Periphery Structure

Moving from frequency to connectivity, the next analysis examines how individual nodes connect to other nodes within the same discipline. For each node, the internal strength s_{int} is computed as the sum of the weighted links connecting it to other nodes in its own discipline. The distribution of these values across all 3,585 nodes is shown in

Figure 22 (left panel), and it is well described by the sum of two Gaussian functions.[26] The means of these two components fall at approximately 25 and 140 on the strength scale, suggesting that nodes within a cluster divide into two structurally distinct groups.

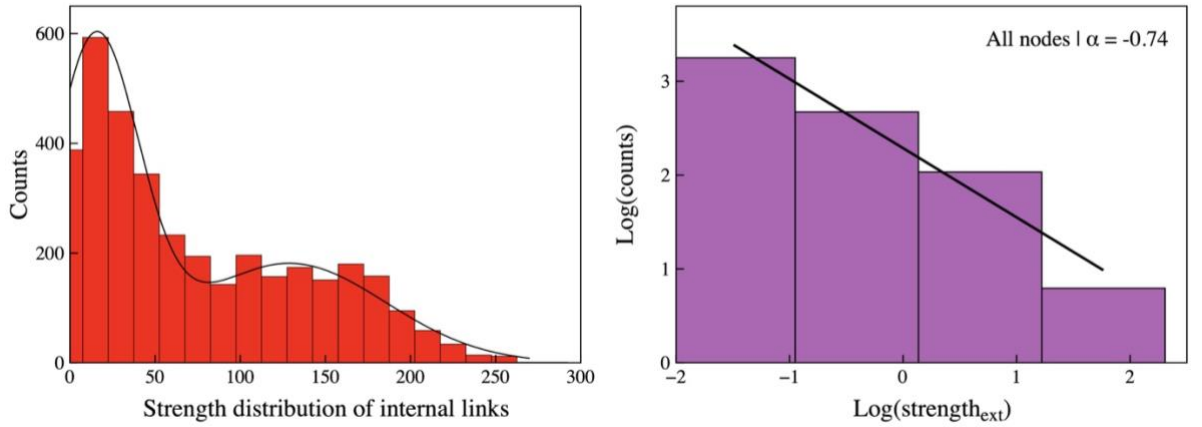


Figure 22 Distributions of the average link strength for all nodes in the network. (**Left panel**) Distribution of internal link strength, showing a bimodal structure captured by the sum of two Gaussian functions (solid curve), which serves as a guide to the eye. (**Right panel**) Distribution of external link strength plotted on a double-logarithmic scale with logarithmic binning, revealing a power-law behaviour with exponent $\alpha = -0.74$ (solid line). The contrasting shapes of the two distributions suggest that internal and external connectivity follow fundamentally different statistical laws.

This bimodal pattern is suggestive of what network science calls a core-periphery structure [34, 35]. A core-periphery network contains a dense, highly interconnected group of nodes (the core) and a sparser, more loosely connected group of nodes (the periphery). The first Gaussian corresponds to the peripheral nodes, which have relatively few intra-cluster connections, while the second corresponds to core nodes, which are densely connected within the cluster. The same bimodal pattern is observed for each of the three disciplinary clusters individually, which means that the core-periphery organization is a general property of intra-cluster connectivity in the 17th century network rather than a feature of any specific discipline.

[26] A Gaussian function is a smooth, continuous mathematical function that produced a symmetrical, bell-shaped curve [33].

This finding indicates that the cultural relations among nodes within a discipline follow a structured organization in which a small group of central entities anchors a much larger group of peripheral ones. In humanistic terms, this is the quantitative counterpart of the well-recognised distinction between canonical and marginal figures within a discipline, but the method used here specifies that distinction structurally rather than by appeal to received reputation. Which nodes specifically occupy the core in each cluster is a question that requires further analysis and is reserved for future work.

Inter-Cluster Connectivity: A Power Law and the Generalist-Specialist Distinction

After examining the connection within the discipline itself, the final node-level analysis examines how individual nodes connect to nodes in other disciplines. The external strength s_{out} is computed analogously to s_{int}, but summing weighted links to nodes outside the cluster. The distribution of s_{out} values, shown in Figure 22 (right), is well described by a power law with an exponent $\alpha = -0.74$. The same general pattern holds when the distribution is computed separately for each discipline.

A power-law distribution of external strength has a specific structural meaning. Most nodes connect only weakly across discipline boundaries, but a small number of nodes connect very strongly (refer to Burt and Soda broker's concept above). The scale of the distribution shows that the ratio between the two groups can be as large as 200 to 1, meaning that for every node that bridges strongly across disciplines, there are roughly two hundred that remain disciplinarily contained. The strongly bridging nodes are called "generalists" and the weakly bridging nodes "specialists". The exponent of the power law quantifies the balance between them: the more negative the exponent, the higher the proportion of specialists to generalists.

This finding is analytically significant for two reasons. First, it offers a quantitative formulation of a distinction that humanistic scholarship has long made qualitatively, namely the distinction between figures who worked across disciplines and figures who

work only within one discipline. Second, the power-law form of the distribution indicates that intra-cluster and inter-cluster connectivity are governed by different generative processes. The bimodal Gaussian distribution of intra-cluster strength is consistent with a random-growth process of the kind first described by Erdős and Rényi [36, 37], in which connections form without systematic preference. By contrast, the power-law distribution of inter-cluster strength is consistent with a preferential-attachment process of the kind described by Barabási and Albert [38, 39], in which already well-connected nodes are more likely to gain further connections. The cultural network therefore appears to grow through two distinct mechanisms operating at different structural levels: random within disciplines, preferential across them. Confirming the precise generative dynamics behind this observation would require further work, but the empirical signature is consistent with the predictions of these two models.

3.1.3. Discussion

Summary of the findings. The analysis presented in this chapter applied a triad-averaging framework to English Wikipedia's wikilinks network in order to characterise the cultural network of Art, Science, and Philosophy in 17th century Europe. The results can be summarised at three analytical scales, moving from the network as a whole to its individual nodes.

At the level of the whole network, the average normalised modularity stabilised at $\overline{Q_N} = 0.862 \pm 0.005$ with an estimation error of 0.6% when all 465 triads are included. This value indicates that the 17th-century cultural network divides quite well into three disciplinary clusters and that this clusterisation is a property of the period rather than an artefact of any particular triad. Placed alongside prior single-triad studies of the Italian Renaissance ($Q_N = 0.77$) [20], and the early 20th century ($Q_N = 0.88$) [19], the 17th-century value sits in the middle of the sequence. The three data points are too

few to confirm a trend, but their ordering is consistent with the historical hypothesis that disciplinary specialisation has increased across the modern period.

At the cluster level, the assortativity matrix revealed two findings. First, Philosophy emerged as the discipline with the most intra-cluster connections ($\overline{\alpha_{33}^{N}} = 0.313 \pm 0.005$), a result that departs from the pattern observed in the Italian Renaissance and the early 20th century, where Art occupied that position. Second, the Science-Philosophy connection ($\overline{\alpha_{23}^{N}} = 0.092 \pm 0.003$) was substantially stronger than the Art-Science ($\overline{\alpha_{12}^{N}} = 0.025 \pm 0.002$) and Art-Philosophy ($\overline{\alpha_{13}^{N}} = 0.017 \pm 0.001$) connections, exceeding them by factors of roughly four and five respectively. The relative variability of these coefficients was also informative: the Science-Philosophy connection was the most homogeneous across triads, while the Art-related connections showed substantially higher variability, with relative standard deviations of 117% for Art-Philosophy and 140% for Art-Science.

The openness measure complemented these findings. Art had an average openness approximately three times lower than that of Science and Philosophy, confirming its peripheral position within the period's cultural network. The triad-level analysis showed that this openness varied systematically with the choice of artist, with English-connected artists such as Peter Lely, Godfrey Kneller, and Christopher Wren producing markedly higher openness values than continental artists.

At the node level, the analysis of 3,585 unique cultural entities yielded two distinct statistical patterns. The distribution of the intra-cluster strength followed a bimodal shape well described by the sum of two Gaussian functions, which is consistent with a core-periphery organisation. By contrast, the distribution of the inter-cluster strength followed a power law with exponent $\alpha = -0.74$, indicating that a small number of nodes carry a disproportionate share of cross-disciplinary connections. The most frequent cultural entities across the 465 triads were dominated by figures and concepts associated with Science and Philosophy (for instance, Pierre Gassendi, Robert Boyle,

"Mechanical philosophy", and "Rationalism"), and most of these high-frequency nodes were not themselves seed entities.

Theoretical Implications. Three theoretical implications follow from the results presented in this chapter, each extending beyond the specific case of the seventeenth century.

The first concerns the analytical status of conceptual nodes within cultural networks. The theoretical literature distinguishes cultural networks from social networks on the grounds that the former include non-human entities such as ideas, movements, and material objects alongside people (see *section 3.1.1. Background*). The results reported here give that distinction an empirical correlate: abstract concepts such as "Mechanical philosophy" and "Rationalism" emerged among the most structurally central nodes in the network without having been selected as seeds, and they did so with frequencies that exceed those of many named individuals. This implies that the structural prominence of conceptual nodes is not merely a theoretical postulate but a measurable, period-specific property. Mapping which kinds of entities occupy central positions in a given period's network could therefore serve as a comparative indicator of how knowledge is organised and transmitted across different historical contexts (see *Results - Embeddings Analysis*).

The second implication concerns the generative mechanisms underlying cultural network formation. The coexistence of two statistically distinct distributions points to the possibility that cultural networks grow through two mechanisms operating simultaneously at different structural scales. A random-growth process, of the kind described by Erdős and Rényi, would generate the bimodal pattern observed within disciplines. A preferential-attachment process, of the kind described by Barabási and Albert, would generate the power-law pattern observed across disciplines. If this dual-mechanism hypothesis can be confirmed through explicit generative modelling, which lies beyond the scope of the present work, it would have substantive consequences

for theories of cultural change. It would suggest that innovation within a discipline and the emergence of cross-disciplinary bridges are not governed by the same social logic, and that the conditions facilitating one need not facilitate the other.

The third implication concerns the relationship between the present findings and existing frameworks in network sociology. The overall geometry of the seventeenth-century network is broadly compatible with Granovetter's weak-ties thesis. However, the power-law tail of the external strength distribution introduces an important qualification. If cross-disciplinary connectivity were uniformly distributed across weakly bridging nodes, as a straightforward application of Granovetter's framework would suggest, then the distribution would not exhibit such a pronounced tail. The fact that it does indicates that a small number of nodes carry a share of integrative capacity, which corresponds more closely to Burt and Soda's model of brokers spanning structural holes. The seventeenth-century cultural network therefore exhibits a hybrid bridging regime that neither framework fully captures on its own. This suggests that future theoretical work on cultural networks may need to combine elements of both accounts, reserving the weak-ties framework for the bulk of sparse cross-discipline links while invoking a brokerage logic for the concentrated tail. More broadly, it raises the question of whether this hybrid structure is particular to the seventeenth century or a general feature of cultural networks. Such type of question can be answered by the comparative framework introduced here once more periods are analysed.

Finally, the period-averaging procedure introduced in this chapter has a methodological implication that extends to other historical contexts. By averaging across many triads sampled from the same period, the analysis recovers structural properties of the period itself rather than properties of any particular configuration of individuals within it. This shifts the unit of analysis from the individual triad to the historical period, and it provides a framework within which different periods can be compared on the same statistical footing. Although only three periods (Italian Renaissance, seventeenth century, early twentieth century) can currently be placed on

this scale, the framework establishes the conditions under which more systematic cross-period comparison becomes possible.

Practical Implications. The findings also carry implications for the practice of cultural and historical analysis, particularly for scholars working at the intersection of digital humanities, cultural analytics, and complex network theory. For digital humanists working with Wikipedia or comparable knowledge graphs, the results indicate that the wikilinks network can be treated as a meaningful empirical object rather than as a mere accessory to article content. The agreement between the structural patterns recovered here and well-established historical knowledge supports the practical use of wikilinks networks as a data source for cultural research. This is methodologically significant because wikilinks are less affected by the textual biases of Wikipedia than article content itself. Editorial prejudices and partial framings tend to reside in the wording of articles, while the linking structure aggregates many small editorial decisions. The use of the normalised Google distance, which relies on multiple links rather than direct connections, further reduces the influence of any single editorial choice on the resulting network.

For historians working on specific periods, the framework offers a tool for identifying cultural entities that merit closer examination. The fact that most of the high-frequency nodes in the seventeenth-century network were not themselves seed entities means that the method can surface figures and concepts whose structural importance is not obvious from a direct reading of the primary sources. Pierre Gassendi, Pierre Bayle, and "Mechanical philosophy" emerge as central not because they were selected at the outset but because they occupy structurally important positions in the network of relations that the seed entities collectively generate. This is the analytical capability that the focus-reading framework introduced in the *Introduction* is designed to support: large-scale pattern detection identifies what is worth examining closely, and close reading then takes over.

For scholars interested in the relationship between disciplines, the method provides a way of quantifying distinctions that have previously been articulated qualitatively. The distinction between "generalists" and "specialists," for instance, has long been part of humanistic discourse about intellectual history, but the power-law form of the external strength distribution allows this distinction to be measured and compared across periods. Similarly, the distinction between core and peripheral figures within a discipline, which is often inherited from canon formation, can be specified structurally rather than by appeal to received reputation. These quantitative formulations are not intended to replace qualitative judgement, but to provide an additional layer of evidence that can be brought into dialogue with humanistic interpretation.

Bibliography

1. Miccio LA, Agapitos P, Gamez-Perez C, et al (2025) Wikipedia as a cultural lens: a quantitative approach for exploring cultural networks. Humanit Soc Sci Commun 12:462. https://doi.org/10.1057/s41599-025-04772-5

2. Schich M, Song C, Ahn Y-Y, et al (2014) A network framework of cultural history. Science 345:558–562. https://doi.org/10.1126/science.1240064

3. Michel J-B, Shen YK, Aiden AP, et al (2011) Quantitative Analysis of Culture Using Millions of Digitized Books. Science 331:176–182. https://doi.org/10.1126/science.1199644

4. Gao J, Hu J, Mao X, Perc M (2012) Culturomics meets random fractal theory: insights into long-range correlations of social and natural phenomena over the past two centuries. J R Soc Interface 9:1956–1964. https://doi.org/10.1098/rsif.2011.0846

5. Turchin P (2008) Arise “cliodynamics.” Nature 454:34–35. https://doi.org/10.1038/454034a

6. Smolla M, Akçay E (2019) Cultural selection shapes network structure. Sci Adv 5:eaaw0609. https://doi.org/10.1126/sciadv.aaw0609

7. Brown DM, Soto-Corominas A, Suárez JL (2016) The preliminaries project: Geography, networks, and publication in the Spanish Golden Age. Digital Scholarship Humanities fqw036. https://doi.org/10.1093/llc/fqw036

8. Suarez J-L, McArthur B, Soto-Corominas A (2015) Cultural Networks and the Future of Cultural Analytics. In: 2015 International Conference on Culture and Computing (Culture Computing). IEEE, Kyoto, Japan, pp 95–98

9. Ahnert R, Ahnert SE, Coleman CN, Weingart SB (2020) The Network Turn: Changing Perspectives in the Humanities, 1st ed. Cambridge University Press

10. Mehdi M, Okoli C, Mesgari M, et al (2017) Excavating the mother lode of human-generated text: A systematic review of research that uses the wikipedia corpus. Information Processing & Management 53:505–529. https://doi.org/10.1016/j.ipm.2016.07.003

11. Rantala H, Leskinen P, Peura L, Hyvönen E (2024) Representing and Searching Associations in Cultural Heritage Knowledge Graphs Using Faceted Search. In: Salatino A, Alam M, Ongenae F, et al (eds) Studies on the Semantic Web. IOS Press

12. Fischer F, Blakesley J, Wojcik P, Jäschke R (2023) Preface: World Literature in an Expanding Digital Space. jca 8:1305. https://doi.org/10.22148/001c.74598

13. Zheng X, Chen J, Yan E, Ni C (2023) Gender and country biases in Wikipedia citations to scholarly publications. Asso for Info Science & Tech 74:219–233. https://doi.org/10.1002/asi.24723

14. Eom Y-H, Shepelyansky DL (2013) Highlighting entanglement of cultures via ranking of multilingual Wikipedia articles. PLOS ONE 8:1–10. https://doi.org/10.1371/journal.pone.0074554

15. Eom Y-H, Aragón P, Laniado D, et al (2015) Interactions of Cultures and Top People of Wikipedia from Ranking of 24 Language Editions. PLOS ONE 10:e0114825. https://doi.org/10.1371/journal.pone.0114825

16. Arroyo-Machado W, Torres-Salinas D, Herrera-Viedma E, Romero-Frías E (2020) Science through Wikipedia: A novel representation of open knowledge through co-citation networks. PLoS ONE 15:e0228713. https://doi.org/10.1371/journal.pone.0228713

17. Gabella M (2019) Cultural Structures of Knowledge from Wikipedia Networks of First Links. IEEE Trans Netw Sci Eng 6:249–252. https://doi.org/10.1109/TNSE.2018.2812788

18. Ibrahim M, Danforth CM, Dodds PS (2017) Connecting every bit of knowledge: The structure of Wikipedia's First Link Network. Journal of Computational Science 19:21–30. https://doi.org/10.1016/j.jocs.2016.12.001

19. Schwartz GA (2021) Complex networks reveal emergent interdisciplinary knowledge in Wikipedia. Humanit Soc Sci Commun 8:1–6. https://doi.org/10.1057/s41599-021-00801-1

20. Miccio LA, Gámez-Pérez C, Suárez JL, Schwartz GA (2022) Mapping the Networked Context of Copernicus, Michelangelo, and Della Mirandola in Wikipedia. Advs Complex Syst 25:2240010. https://doi.org/10.1142/S0219525922400100

21. Gatti H (2025) Divisions of Knowledge in the European Renaissance: A Case Study. History of Humanities 10:291–309. https://doi.org/10.1086/737034

22. Sandoz R (2021) Thematic Reclassifications and Emerging Sciences. J Gen Philos Sci 52:63–85. https://doi.org/10.1007/s10838-020-09526-2

23. Hobbes T (2017) Leviathan. Penguin Books, Harmondsworth, Meddlesex

24. Blair AM (2007) Organizations of knowledge. In: Hankins J (ed) The Cambridge Companion to Renaissance Philosophy. Cambridge University Press, Cambridge, pp 287–303

25. Burke P (2017) Orders of Knowledge in Early Modern Europe. Asiatische Studien - Études Asiatiques 71:993–1002. https://doi.org/10.1515/asia-2017-0019

26. Wade D (2017) Geometry & art: how mathematics transformed art during the Renaissance. Shelter Harbor Press, New York

27. Suárez JL, Sancho F, De La Rosa J (2012) Sustaining a Global Community: Art and Religion in the Network of Baroque Hispanic-American Paintings. Leonardo 45:281–281. https://doi.org/10.1162/LEON_a_00374

28. Tucker R (2010) The Patronage of Rembrandt's Passion Series: Art, Politics, and Princely Display at the Court of Orange in the Seventeenth Century. The Seventeenth Century 25:75–116. https://doi.org/10.1080/0268117X.2010.10555640

29. Granovetter MS (1973) The Strength of Weak Ties. American Journal of Sociology 78:1360–1380. https://doi.org/10.1086/225469

30. Burt RS, Soda G (2021) Network Capabilities: Brokerage as a Bridge Between Network Theory and the Resource-Based View of the Firm. Journal of Management 47:1698–1719. https://doi.org/10.1177/0149206320988764

31. Casini P (1988) Newton's *Principia* and the philosophers of the enlightenment. Notes and Records of the Royal Society of London 42:35–52. https://doi.org/10.1098/rsnr.1988.0006

32. Boran E, Feingold M (2017) Reading Newton in early modern Europe. Brill, Leiden ; Boston

33. Squires GL (2001) Practical Physics, 4th ed. Cambridge University Press

34. Barucca P, Tantari D, Lillo F (2016) Centrality metrics and localization in core-periphery networks. J Stat Mech 2016:023401. https://doi.org/10.1088/1742-5468/2016/02/023401

35. Gurugubelli S, Chepuri SP (2022) Generative Models and Learning Algorithms for Core-Periphery Structured Graphs

36. Erdős P, Rényi A (1960) On the evolution of random graphs. Publ Math Inst Hungar Acad Sci 5:17–61

37. Barabási A-L (2016) Chapter 3 - Random Networks. In: Network Science, 1st ed. Cambridge University Press

38. Barabási A-L, Albert R (1999) Emergence of Scaling in Random Networks. Science 286:509–512. https://doi.org/10.1126/science.286.5439.509

39. Barabási A-L (2013) Network science. Phil Trans R Soc A 371:20120375. https://doi.org/10.1098/rsta.2012.0375

3.2. Network and Semantic Patterns in Wikipedia's Network of Biographies

3.2.1. Background

The previous section (*Results – section 3.1.2*), using complex network methods, characterised the cultural network of Art, Science, and Philosophy within the 17th century alone. The present section complements this approach to a substantially broader scope, both temporal and disciplinary. Two methodological choices make this complementary step possible. First, WikiTextGraph (see *Methodology - WikiTextGraph*) parses and processes Wikipedia data at scale. Second, a series of filtering steps yields a dataset of approximately 1.9 million biographical pages, classified into Fields of Human Activity (FHA) (see *Methodology - Biographical Network Construction*). This section first reviews the relevant literature and then identifies the research gap that the present analysis addresses.

A growing body of work applies network-based methods to cultural and historical phenomena at scale, and Wikipedia has emerged as one of the most frequently examined repositories within this tradition. Early studies in this area used network centrality metrics on biographical articles to identify the most prominent historical figures across language editions. Their findings reveal a consistent pattern: locally celebrated individuals coexist with a smaller set of globally recognised figures [1–4]. This pattern points to the layered character of cultural prominence, in which local and global recognition operate as distinct but coexisting dimensions. Within this first body of work, disciplinary case studies of mathematicians [5] and philosophers [6] further show that structural indicators of importance, such as in-degree or PageRank scores, remain largely stable over time. Combining disciplinary focus with temporal dynamics, a smaller set of studies documents recency biases and temporal homophily but also some shifts in historical prominence over short periods [7, 8]. More recent work extends this perspective by examining cross-field connections in the 17th century and

identifying individuals who bridge distinct intellectual areas [9–11]. Collectively, these studies establish that network-based approaches are well suited to mapping the relational structure of cultural networks, and of Wikipedia in particular. However, two limitations emerge: the temporal window of analysis is typically narrow, and the focus tends to be restricted to specific disciplines. The analysis presented here addresses both limitations by extending the temporal window to 25 centuries, from the 5th century BCE to the 20th century CE, and by covering multiple FHAs.

A limitation of a different sort, shared by the network-based studies reviewed above is that they treat connections between cultural entities as structural, defined by their position within the network topology. However, the meaning of a link between two biographical figures also depends on the conceptual context in which that link appears. A node connected to many others occupies a structurally central position, but this alone does not reveal whether those connections reflect shared intellectual tradition, institutional proximity, or disciplinary overlap. Recovering that distinction requires attending to the semantic content of the links, not only their topology. A representative example of this integrated approach is the work of Sangiacomo et al. [12], who reconstruct the network of early modern natural philosophy by combining vectorised textual representations with network techniques, thereby coupling the structural and semantic dimensions of intellectual exchange. The present analysis adopts a similar framework. As detailed in the *Methodology*, the second part of the analysis applies semantic embeddings to the wikilinks that compose the biographical network. This combined approach allows the structural patterns identified through network analysis to be examined alongside the semantic contexts in which they are embedded, rather than treating the two points of view as separate phenomena.

The use of embeddings to analyse cultural data connects to a broader methodological tradition known as cultural cartography. Within this tradition, word embeddings are used to construct a high-dimensional mathematical space in which the meaning of a word or concept is defined by its position relative to all other words [13]. To put this in

simple terms: words that appear in similar contexts across large bodies of text end up positioned close together in this space, while conceptually distant words are placed far apart. By projecting vast amounts of text into this geometric representation, cultural cartography translates complex cultural relationships into a form that can be systematically measured and compared. Lee and Martin [14] characterise the resulting representations as "impoverished but intersubjectively valid", meaning that while some nuance is inevitably lost in the translation, the simplified map remains meaningful across different observers and interpretive contexts. Beyond static mapping, a second strand of this literature examines how the positions of words and concepts shift over time. Hamilton et al. [15] demonstrate that, while words carrying cultural markers of social status were in constant flux throughout the 20th century, the underlying dimensional structure of class remained stable. This finding is important because it shows that surface-level semantic change does not necessarily indicate structural transformation. Further work in this strand has tracked historical shifts in gender and ethnic stereotypes [16], revealing how culturally embedded associations evolve across periods.

The studies reviewed above that use embeddings to map cultural phenomena, including the connectivity patterns of biographical networks, predominantly rely on static word embeddings. In this approach, each word or concept is assigned a single fixed vector regardless of the context in which it appears. This works well for capturing broad semantic regularities, but it cannot distinguish between different uses of the same term. The word "revolution," for instance, carries a different meaning in a biography of a political leader than in the biography of a physicist, yet a static embedding assigns both occurrences the same position in semantic space. The present study makes use of contextual embeddings to embed outgoing links of Wikipedia biographical pages, which generate a distinct vector for each occurrence of a link depending on its surrounding links. The use of contextual embeddings has been shown to capture finer-grained semantic distinctions than static representations [17–

19]. Adopting contextual embeddings within a cultural cartography framework extends the existing literature in two directions. First, it preserves the macroscopic, map-like perspective characteristic of this tradition, allowing large-scale cultural patterns to be observed across centuries. Second, it improves the semantic resolution at which those patterns can be examined, making it possible to detect more subtle shifts in how concepts are used across time and discipline. Building on this improved resolution, the analysis focuses on FHAs associated with knowledge production, particularly Science and Philosophy, in order to trace how their conceptual connections to other fields of human activity change across 25 centuries.

3.2.2. Results

3.2.2.1. Network Analysis

This subsection examines a biography-to-biography network in which nodes represent aggregated biographical pages. Each node combines an FHA with a century, meaning that the network operates at the level of FHA-temporal clusters rather than that of discrete persons. Edges between nodes were normalised through two complementary methods: an outgoing-based normalisation, which accounts for the referential weight of the source node, and an incoming-based normalisation, which reflects the receptive weight of the target node. Together these two perspectives allow connectivity patterns to be read from the perspective of influence and that of reception, as explained in the Methodology section (see *2.5.4. Weight Normalisation & Network Filtering*).

Following the analytical sequence established in the previous section (see *Results – section 3.1*), this section begins at the most aggregated level of observation: the temporal structure of the biographical network as a whole. Figure 23 presents a kernel density estimation (KDE) of links across centuries for all FHAs. A KDE plot is suitable for big data visualisation because it shows the shape of the data's distribution as a smooth curve, making it easy to see where values are concentrated and where they

are sparse [20]. The plot shows a relatively strong and constant (without gaps) concentration of connections along the diagonal. This means that biographical articles predominantly link to other biographies of individuals born within the same or immediately adjacent centuries. Quantitatively, 96.99% of all biographical links occur within a one-century span, a level of temporal concentration not previously demonstrated at this scale. Since the century bin is coarse and it can conflate distances up to ~99 years, we move to the level of years. At this level we see that the average temporal distance between linked individuals is 33.78 years, a value that is of the same order as the average human generational interval of 27 to 30 years reported in demographic research [21, 22].

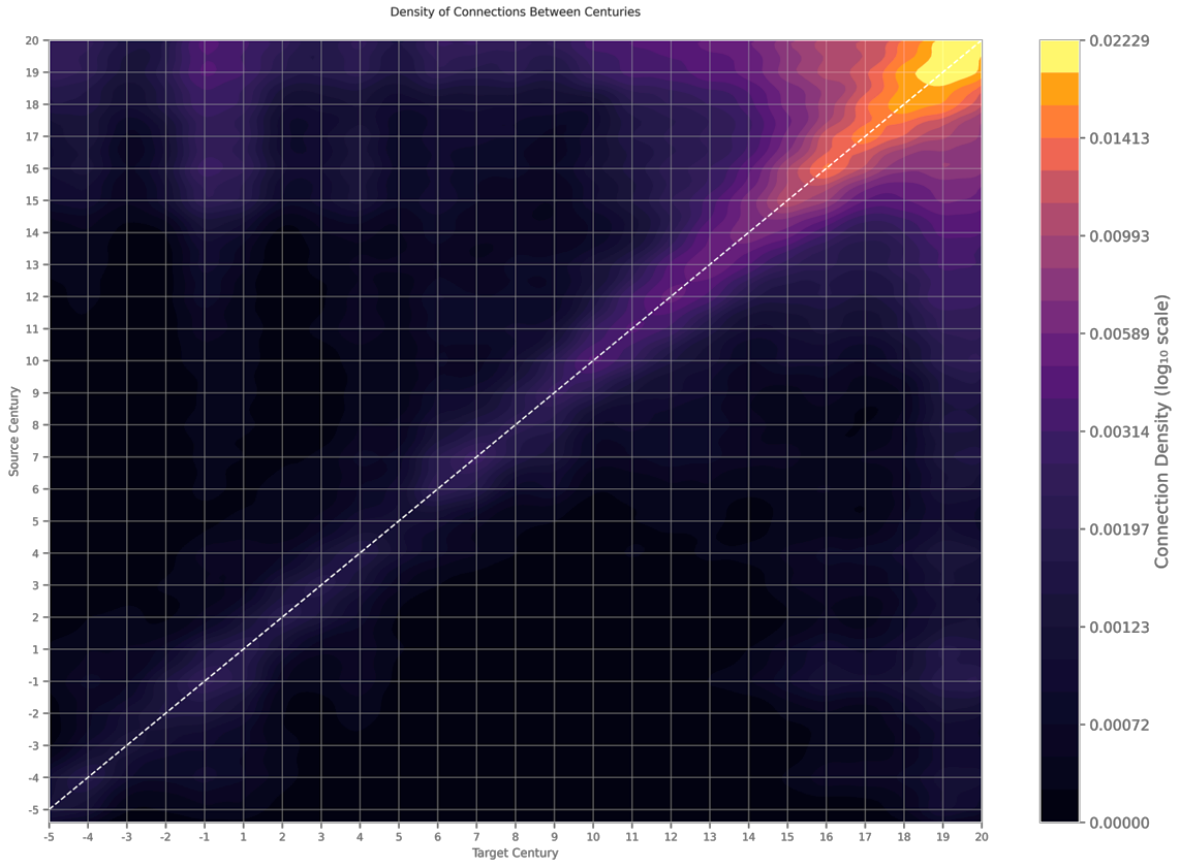


Figure 23 Cross-temporal connection density between biographical subjects across source and target centuries ($\log_{10}$ scale).

This finding shows a pattern of temporal homophily, defined here as the tendency of biographical pages to connect preferentially to individuals born in temporally

proximate periods. Moreover, the fact that this result appears when considering a large number of biographies indicates that it is a structural property of Wikipedia's biographical network. While Jatowt *et al*. [7] previously identified this pattern, their analysis was limited to a narrower time frame; the present study extends it across 25 centuries and nearly two million biographies, demonstrating that temporal homophily is not a period-specific feature but a general organising principle of the network.

Two further observations can be made for this temporal homophily pattern. The first one is regarding the 33.78-year average. The metric is computed across all FHAs and the direction in this case does not matter. That means that the result is an aggregation of a heterogeneous set of relations, which can include kinship and dynastic ties, marriage, succession in office, collaboration, rivalry, sporting partnership, etc. [23]. Each of these type of relations can independently connect individuals whose active lives overlap within roughly one generation. What the convergence with the demographic interval suggests is that the link structure of Wikipedia registers the temporal radius of social co-existence as refracted through the lens of encyclopaedic documentation.

The second observation qualifies the strength of the diagonal pattern. While 96.99% of biographical links fall within a one-century span, this near-total temporal clustering does not mean that long-range connections are absent. Such connections are statistically rare; however, as the subsections that follow demonstrate, they are not randomly distributed. Lastly, on the diagonal we observe that the link density increases from the 15th century onward, reflecting both the expansion of Wikipedia's biographical coverage in more recent periods and the well-documented asymmetries in how the online encyclopaedia represents the recent past relative to more distant eras.

Temporal Homophily within Specific Disciplinary Pairings

Within Science, 96.02% of biographical links fall within a one-century span, close to the full-network figure of 96.99%. This proportion declines once a link crosses a disciplinary boundary: 85.84% for links running from Humanities & Philosophy to Science, and 91.46% in the reverse direction. Based on these numerical insights, there are three points that worth mentioning.

First, temporal homophily persists across disciplinary boundaries, but it weakens in that context, so cross-field connections are where the network's temporal coupling relaxes most within this comparison. The relaxation is also asymmetric: links originating in Humanities & Philosophy are more temporally dispersed than those originating in Science (85.84% within one century, against 91.46% in the reverse direction).

This dispersion is visible in the KDE plots (see **Figures 24A-B**). More specifically, figure 24A shows diffuse off-diagonal density, including a pronounced cluster in the upper-left region where recent figures (source ≈ 19th–20th century) link back to Classical-era science (target ≈ 5th–3rd century BCE). One plausible reading of that "hotspot", which requires node-level inspection to confirm, is that modern philosophers of science account for a substantial share of these links. On the other hand, the plot for Science to Humanities & Philosophy remains more tightly bound to the diagonal (see **Figure 24B**).

Reading Figure 24B more closely, three hotspots of connections stand out. The first sits in the lower-left corner (source and target both in the Classical era, roughly the 5^{th}-3^{rd} century BCE), where scientists of antiquity direct a large share of their cross-field links to Humanities & Philosophy figures of the same period. The second lies on the diagonal around the 5^{th} century CE, and the third around the 11^{th}-12^{th} century CE. Because the panel is normalised by outgoing links, these cells indicate that scientists that lived in those centuries concentrate a very high proportion of their Humanities-directed links within a narrow contemporaneous band.

These three regions lie on or very close to the diagonal, which is consistent with the temporal-homophily pattern reported above: even when a link cross the disciplinary boundary, we see that it tends to stay within roughly one century. What the hotspots add is that this cross-field coupling is not temporally uniform as it becomes denser in particular eras. With some caution, someone could read these as periods in which the documented activity of scientists and of philosophers was closely intertwined: late antiquity and the medieval period are both eras in which natural philosophy and philosophy were not yet institutionally separated, so an individual identified retrospectively here as "Science" and one identified as "Humanities & Philosophy" may describe figures working within a single intellectual milieu. This reading should be viewed here as a hypothesis that the observed structure is compatible with and not as a claim that the density per se establishes. This is because to confirm that, it would require examining the specific individuals in each cluster.

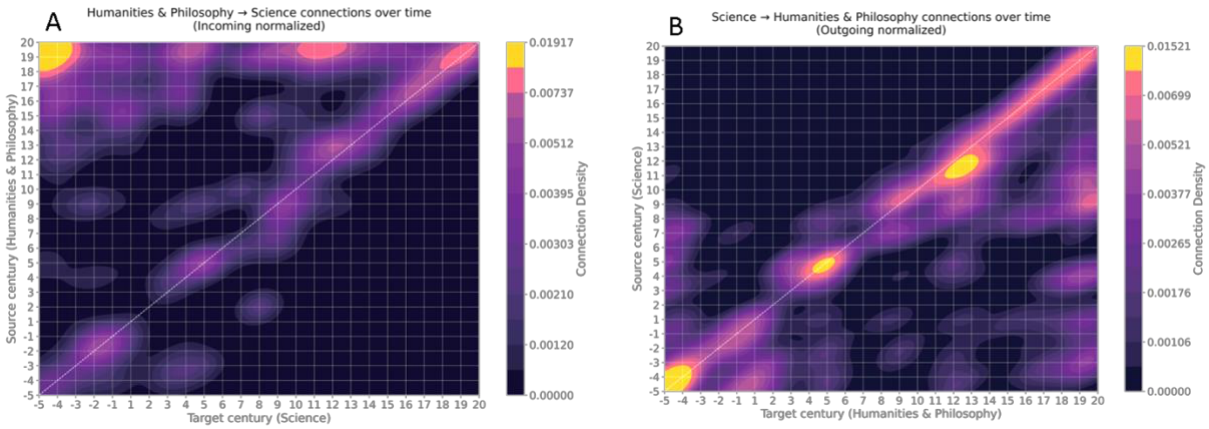


Figure 24 Kernel density estimation (KDE) plot of (**A**) in-degree–normalized cross-field links between Humanities & Philosophy and Science across centuries and (**B**) out-degree normalized between Science and Humanities & Philosophy. Lighter shades indicate higher link density.

Long-Range Connections and the Topological Bypass

The *Introduction* established that Wikipedia's network, as approached in this thesis, constitutes a cultural network [28]. A cultural network, even if the nodes are only biographical pages of people, is not merely a social network. That means that the time in biographical pages is not represented solely as a neutral chronological sequence

but rather as a socially constructed and relational framework that follows the narrative and thus functions across multiple dimensions [29, 30]. That means that, even though biographical pages are anchored to fixed temporal markers such as dates of birth and death, their connection operate through narrative and relational temporalities. That implies that the apparent temporal distance in Wikipedia biographies can be shortened or lengthened depending on the narrative structure, shared FHAs, or perceived intellectual lineage [31, 32],[27] , and it is this narrative-driven connectivity that permits the existence of long-range links, either to the past or to the future. As shown in the preceding section, these long-range links constitute only a small fraction of the total. Nevertheless, they are not randomly distributed. The way in which biographies connect across centuries reveals something distinctive about how specific historical periods are represented in Wikipedia and, by extension, about how the encyclopaedia constructs its version of the past.

Turning to the off-diagonal patterns, the analysis identifies a recurrent feature across all examined pairings of FHAs: a pronounced reduction in link density for biographies spanning the 6th to the 10th centuries CE. In the KDE plots (see **Figures 24, 25, and 26**), this appears as a dark band running through the corresponding region of the matrix. Because the sparsity is visible under both incoming and outgoing normalisation schemes, it cannot be attributed to link directionality alone; rather, it reflects the relative isolation of this period within the network's topology.

[27] The English Wikipedia biography of Isaac Newton exemplifies this temporal expansion and compression. His entry contains outbound links to contemporaries such as Gottfried Wilhelm Leibniz, but also establishes direct connections to individuals from the distant past (e.g., Aristotle) and the Modern Era (e.g., the biographer Richard S. Westfall).

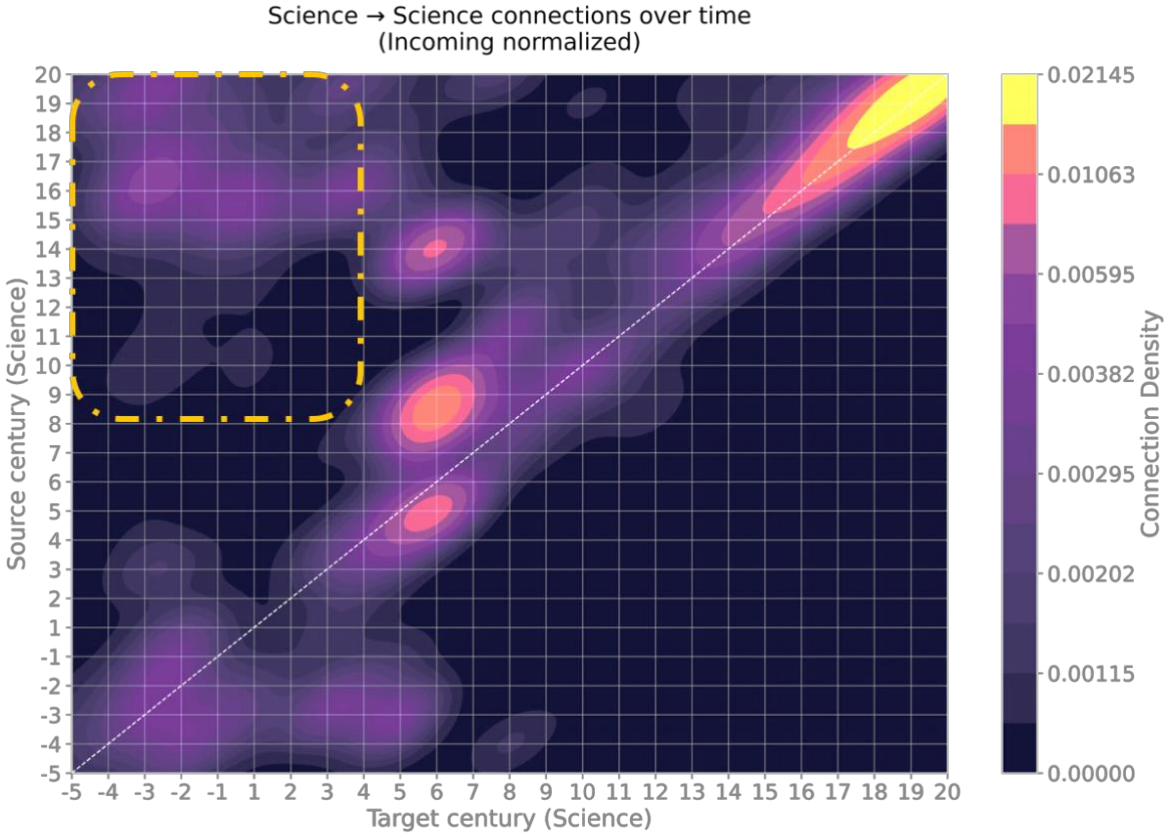


Figure 25 KDE plot of normalized in-degree-based connections from Science (y-axis) to Science (x-axis); this plot illustrates "Out of all the links that Target receives, what is the fraction of links originating from Source?"

To examine this pattern more closely, the analysis focuses on a more specific directional question that spans the "isolated" period identified above: how do scientists from the 8th to the 20th centuries CE link to figures from the Classical Era[28] (5th century BCE to 4th century CE)? Figure 25 shows that retrospective connections to Classical Era scientists appear as early as the 8th century CE, increase markedly during the Renaissance, and then decline modestly in the Modern Era. This trajectory is consistent with what historians of science call the Continuity thesis (i.e., the argument that the intellectual foundations of the Modern Era rested on the sustained

[28] The general historical periods and their time windows used in this study are: Classical Era (5th BCE to 4th century CE), Middle Ages (5th to 14th CE), Early Modern Era (15th to 17th century CE), and Modern Era (18th to 20th century CE). It should be noted that this periodization is used only as a heuristic marker to facilitate the large-scale longitudinal comparison performed here. For this reason, it is important to acknowledge that these boundaries are only historiographical conventions and not absolute temporal containers [33, 34].

transmission and reinterpretation of Classical knowledge, with medieval scholars serving as essential intermediaries [35, 36]. The density of connections visible from the 8th century onwards corresponds to the documented scope of the Islamic Translation Movement, the large-scale effort to translate, adapt, and preserve Greek scientific, philosophical, and medical texts into Arabic, undertaken roughly between the 8th and 10th centuries CE [37–39]. On this reading, the European Renaissance was not the origin of a reconnection with Classical thought; the network renders it instead as the moment when a much longer transmission process became especially concentrated and visible, beginning from the 14th century onward.

Despite this evidence of continuity, a structural asymmetry emerges when the two normalisation perspectives are compared directly. When looking through the lens of the incoming normalisation[29] (see **Figure 25**), medieval scholars appear as connectors; they receive links from Classical Era figures and send links forward to later periods. However, through the lens of the outgoing normalisation[30] (see **Figure 26**), Modern Era scientists tend to link directly to Classical Era figures, bypassing medieval intermediaries. The medieval contribution to the chain of transmission becomes structurally less visible; not because they are absent from the data, but because the outgoing perspective does not bring this contribution to the foreground.

[29] Incoming normalisation asks: of all links a given century receives, what fraction originates from each source century?

[30] Outgoing normalisation asks the reverse: of all links a given century sends out, what fraction goes to each target century?

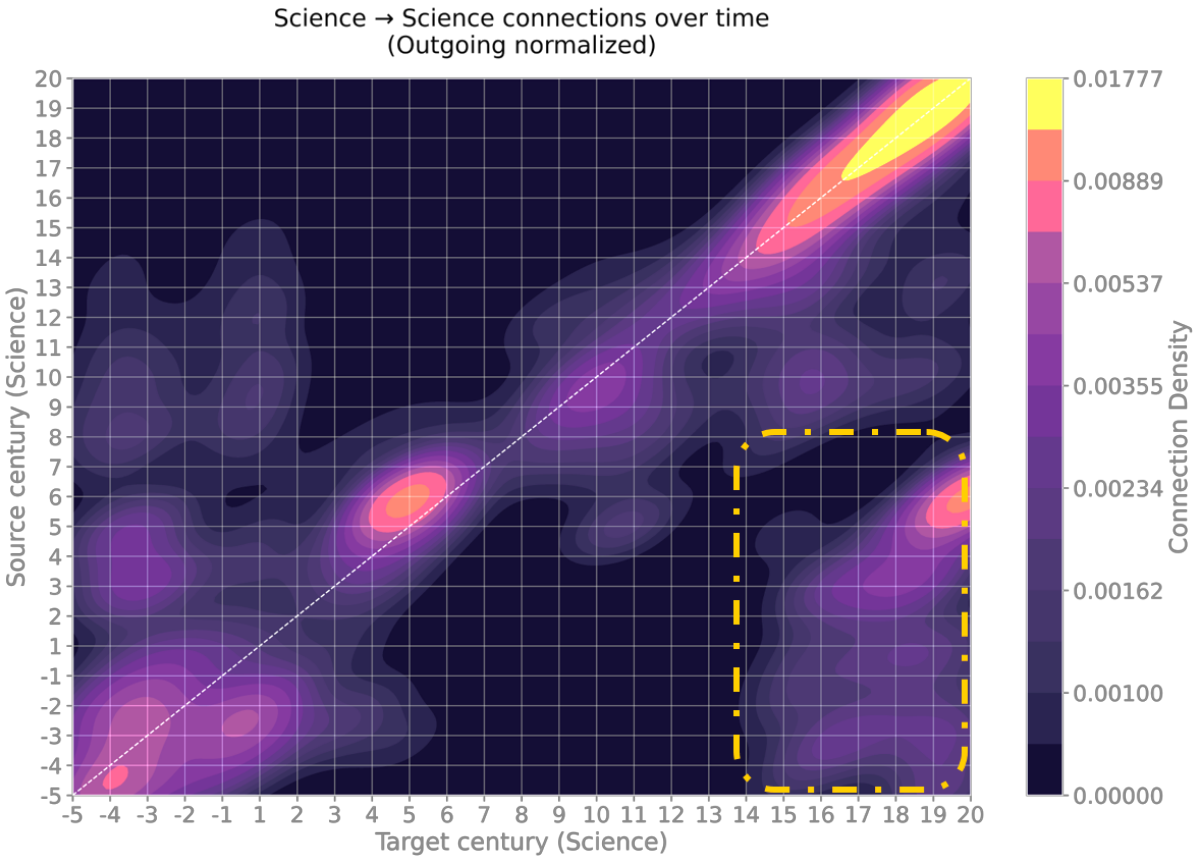


Figure 26 KDE plot of normalized out-degree based connections from Science (y-axis) to Science (x-axis); lighter shades indicate higher link density. This plot can be read as "Out of all the links that cluster Source sends out, what percentage goes to cluster Target?"

This asymmetry carries a specific significance for digital humanities and studies that use computational approaches on knowledge repositories. As presented in the *Introduction*, Wikipedia's link structure does not only reflect historical information, but it also reflects how contemporary editors document and connect this historical information. The pattern visible under outgoing normalisation, in which Modern Era scientists connect directly to Classical antecedents, is consistent with a historiographical narrative in which the Renaissance marks a rupture after a not so active medieval period and a rediscovery of Antiquity through classical texts. This narrative has been substantially revised by recent scholarship [35, 36, 40, 41], yet it remains structurally visible in the network under the outgoing normalisation perspective. This need not mean that individual editors consciously reproduce a superseded view. It is more likely that widely shared cultural framings of intellectual

history leave measurable traces in large-scale link structures as an emergent property of many independent editorial decisions [42, 43]. Lastly, this is the kind of pattern that is difficult to detect through close reading of individual articles but becomes visible at the network scale when data is aggregated.

These results also point to a methodological consideration relevant to anyone working with large digital repositories. As shown here, the choice of normalisation direction is not a neutral technical step; it determines which historical narrative the data makes legible. The same network simultaneously supports a reading of historical continuity, under incoming normalisation, and a reading of historical rupture, under outgoing normalisation. It should be noted that neither reading is more correct than the other, and neither can be derived from the data without a prior interpretive choice about perspective. This is consistent with the argument, developed within digital humanities, that quantitative representations of cultural data always involve decisions that shape what becomes visible [44]. In this thesis, this is not necessarily a limitation but an analytical resource. As described in *Methodology*, incoming and outgoing normalisation are applied in a complementary manner precisely so that the narratives encoded in Wikipedia's biographical network can be examined from both directions.

3.2.2.2. Embeddings Analysis

The preceding section established that biographical connections between FHAs cluster overwhelmingly among contemporaries, with long-range links forming a sparse, non-random layer above this dominant structure. What network topology alone cannot reveal, however, is whether a comparable temporal organisation governs the semantic contexts through which biographical pages engage with the broader world: the concepts, institutions, events, and artefacts that their outgoing links reference. The embeddings analysis addresses this question directly. In keeping with the approach taken above, the focus falls primarily on Science and Humanities & Philosophy, while results across all FHAs combined are also reported, whenever possible.

The analysis extends the embedding approach to the full corpus of Wikipedia biographical articles, each grouped by FHA and century as in the preceding subsection (see *3.2.1.1. Network Analysis*). Unlike the biographical network, however, this analysis does not restrict outgoing links to connections between persons; all outgoing links from each biographical article are retained. The informational content of each FHA-century cluster is then summarised through a centroid vector, computed as the mean of the embedding vectors belonging to that group. This representation condenses the semantic profile of each cluster into a single point in the embedding space, aiming to map the conceptual cultural landscape of each FHA and century as it is embedded in the outgoing links of Wikipedia.

Gradual Temporal Ordering in Semantic Space

The century-level centroid vectors for Science and for Humanities & Philosophy and of all FHAs combined are projected into three dimensions using Principal Component Analysis (PCA), a technique that is suitable for big data visualisation since it reduces the dimensionality of the embedding space while preserving the main axes of variation [20]. This allows the relative positions of different centuries to be visualised and compared.

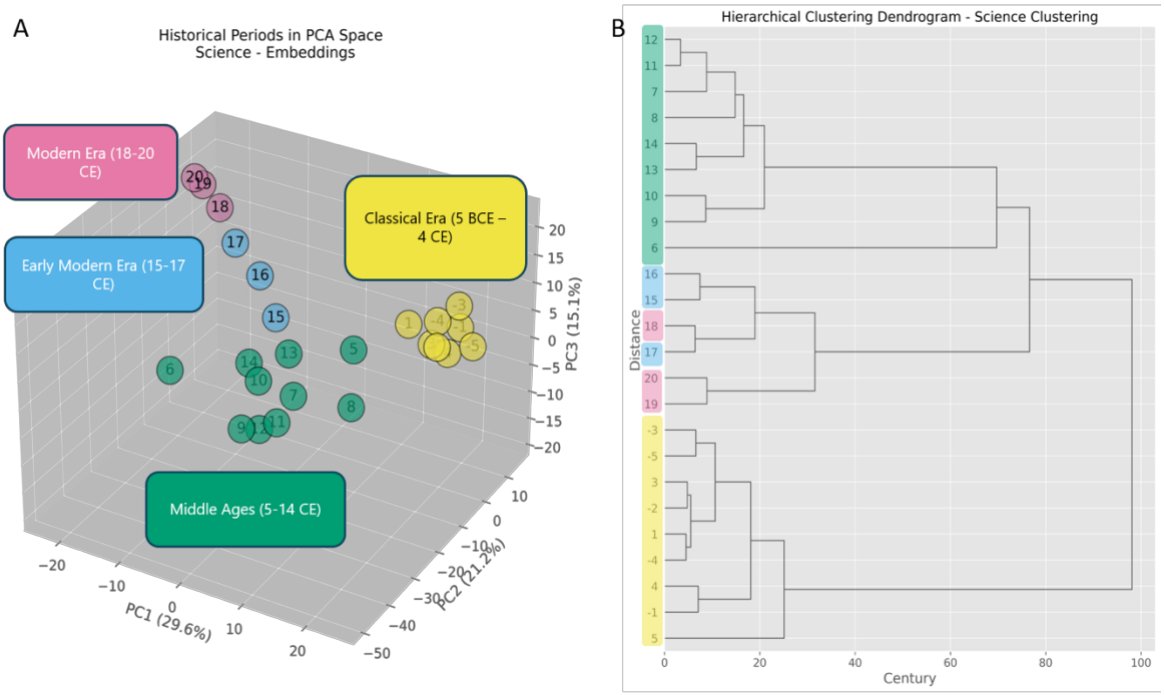


Figure 27 Visual representations of century embeddings for Science. (**A**) A three-dimensional principal component analysis (PCA) projection, where each point denotes a century (negative values indicate BCE, positive values CE) positioned by the first three principal components. Colors represent broad historical periods. (**B**) A hierarchical clustering dendrogram for Science using Ward's linkage and cosine distances, illustrating similarity relationships between centuries without dimensional reduction.

In both projections (see **Figures 27A and 28A**), the century centroids display a tendency toward gradual temporal ordering, particularly over the last eleven centuries, where temporally adjacent centuries tend to occupy relatively close positions in the embedding space. At a broader scale, the centroids also form partially overlapping clusters that correspond closely to conventional historical periods: the Classical Era (5th century BCE to 4th century CE), the Middle Ages (5th to 14th centuries), the Early Modern Era (15th to 17th centuries), and the Modern Era (18th to 20th centuries). It should be mentioned that these groupings are not imposed by the analysis but emerge directly from the embedding structure, without any predefined historical categorisation.

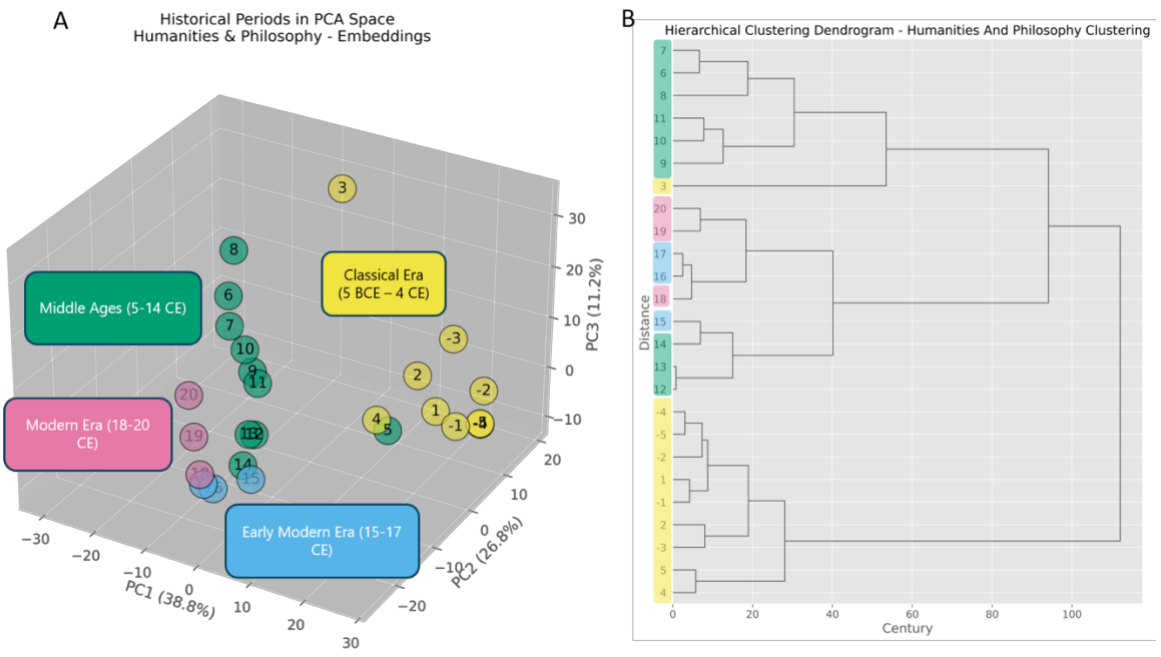


Figure 28 Visual representations of century embeddings for Humanities & Philosophy. (**A**) Three-dimensional principal component analysis (PCA) projection, where each point represents a century (negative values indicate BCE, positive values CE) positioned by the first three principal components. Colors indicate broad historical periods. (**B**) Hierarchical clustering dendrogram for Humanities & Philosophy using Ward's linkage and cosine distances, illustrating similarity relations between centuries without dimensional reduction.

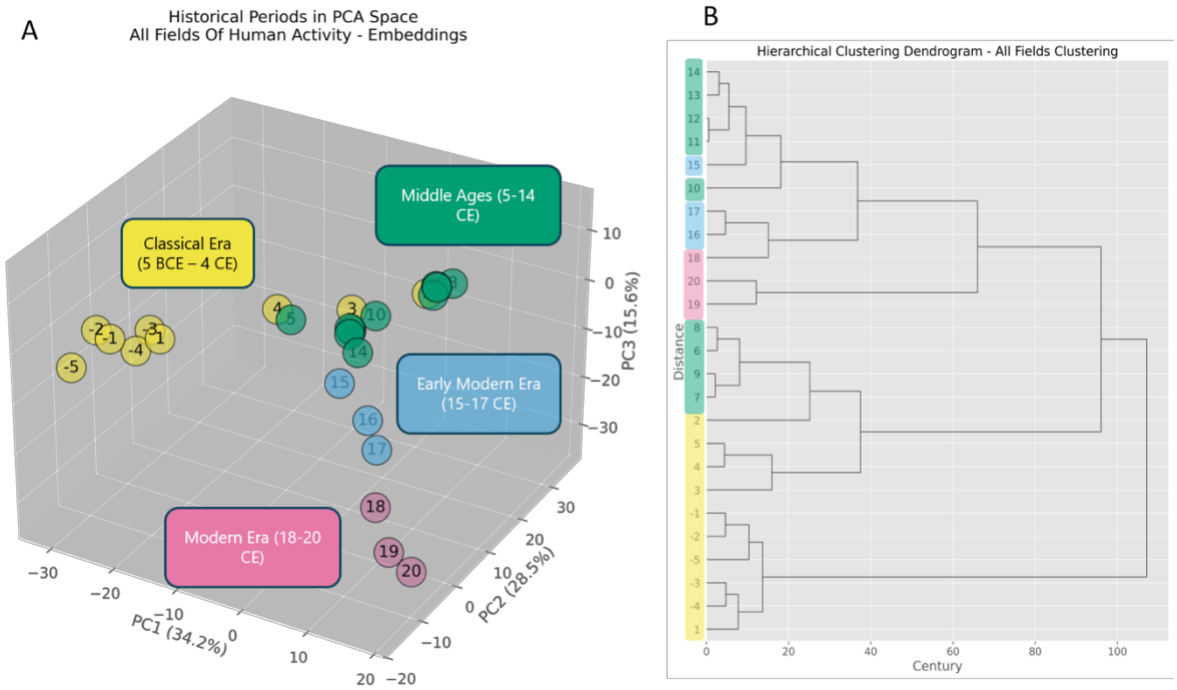


Figure 29 Visual representations of century embeddings for all FHAs combined. (**A**) A three-dimensional principal component analysis (PCA) projection, where each point represents a century (negative values indicate BCE, positive values indicate CE) positioned along the first three principal components. Colors

denote broad historical periods. (**B**) A hierarchical clustering dendrogram for all FHAs combined using Ward's linkage and cosine distances, illustrating similarity relationships between centuries without dimensional reduction.

Because PCA compresses high-dimensional data into a small number of dimensions, there is a risk that the visual closeness of certain centuries in the projected space is a by-product of the compression itself, rather than evidence of genuine similarity in the original embeddings. To rule out this possibility, the temporal patterns observed in the PCA projection are subjected to a series of independent verification steps.

The first verification applies hierarchical clustering, a method that groups data points by progressively merging those that are most similar. Importantly, this procedure is applied directly to the embeddings, not to any reduced-dimensional representation, and it does not impose any prior assumptions about how the centuries should be grouped. The result is shown as a dendrogram, a tree-like diagram in which centuries judged more similar are joined by shorter branches, while more distant ones merge later through taller branches. If the dendrogram recovers recognizable historical periods without being explicitly instructed to do so, this provides independent evidence that the temporal signal is present.

The second verification examines whether the known historical periods (i.e., Classical Era, Middle Ages, Early Modern Era, and Modern Era) correspond to genuinely distinct regions of the embedding space. Two standard measures of cluster quality are used for this purpose. The Silhouette Score captures, for each century, how closely it resembles other centuries belonging to the same period compared to those in the nearest neighbouring period; a high score indicates that the periods are well separated. Additionally, the Calinski-Harabasz Score provides a complementary perspective by comparing the spread of centuries within each period to the spread between periods; again, a higher value reflects cleaner separation.

The third and final verification asks whether the Silhouette Score obtained could simply have arisen by chance, under a null model in which period labels are randomly

reassigned to centuries. To answer this, the period labels are randomly reassigned to centuries a large number of times, and the Silhouette Score is recalculated for each random arrangement. This produces a reference distribution representing what scores would look like if the groupings were meaningless. The actual observed score is then compared against this distribution: if fewer than 5%[31] of the random arrangements produce an equal or higher score, the clustering is considered statistically significant. In practical terms, this means that the organisation of centuries into historical periods is unlikely to be a coincidence and instead reflects structure that is genuinely encoded in the embeddings.

Table 5 presents the cluster quality of historically defined periods against a random baseline across three groupings: Science (abbreviated "Sci.", shown in green), Humanities & Philosophy (abbreviated "Hum. & Phil.", shown in blue), and the full set of FHAs combined (shown in dark pink).

	Sci.	**Hum. & Phil.**	**All FHAs**
Silhouette Score[32]	0.33	0.26	0.22
Mean Null Silhouette Score[33]	-0.28	-0.28	-0.28
P-value	0.0001	0.0001	0.0001

[31] This is also known as the significance threshold usually denoted by α. It is a predetermined probability cutoff used in hypothesis testing to decide whether the results are statistically significant [20].

[32] The Silhouette Score measures how well each data point fits within its assigned cluster relative to neighbouring clusters. For each century, it computes the average dissimilarity to all other centuries in the same period and compares this to the average dissimilarity to centuries in the nearest other period. The resulting value ranges from −1 to 1: a score close to 1 indicates that the century is much more similar to its own period than to any other; a score near 0 suggests it lies on the boundary between two periods; and a negative score indicates it may have been assigned to the wrong group. When averaged across all centuries, the Silhouette Score thus provides a single, interpretable measure of how cleanly the historically defined periods partition the embedding space.

[33] The Mean Null Silhouette Score is the average Silhouette Score obtained by repeatedly reassigning the period labels to centuries at random (in this case across 10,000 permutations) and recalculating the score each time. Because these reassignments sever any connection between a century and its actual historical period, the resulting distribution captures what clustering quality would look like in the absence of any genuine temporal structure. The mean of this null distribution therefore serves as a baseline: a positive observed Silhouette Score that substantially exceeds it provides evidence that the period groupings reflect real organisation in the data rather than a statistical artefact.

The results reported in Table 5 confirm this across all three corpora. The Silhouette Scores are positive in every case — 0.33 for Science, 0.26 for Humanities & Philosophy, and 0.22 for all FHAs combined — indicating that the historically defined periods correspond to genuinely distinct regions of the embedding space. Moreover, when tested against a null model (third verification step), these values stand in contrast to the mean Silhouette Scores obtained from 10,000 random permutations of the period labels, which are negative throughout (−0.28 for all examined cases). A negative null score means that, when centuries are assigned to periods arbitrarily, the resulting groups are less cohesive than what would be produced by a random partition of the space. The permutation test strengthens this: in all three corpora the p-value is 0.0001, indicating that none of the 10,000 random arrangements yielded a Silhouette Score equal to or higher than the one actually observed. Given the aforementioned results, I argue that the probability that the recovered period structure arose by chance is therefore negligible.

Moreover, the dendrograms are consistent with the PCA projections. For both Science (see **Figure 27B**) and Humanities & Philosophy (see **Figure 28B**), as well as for all FHAs combined (see **Figure 29B**), centuries that are close in time tend to merge at lower dissimilarity levels, indicating greater semantic similarity among temporally adjacent periods. At the same time, the larger branches of the dendrograms merge at relatively high dissimilarity levels, reflecting the gradual nature of conceptual change: the semantic profiles of centuries do not sharply divide into distinct historical blocks but shift incrementally over time. This confirms that the temporal signal visible in the PCA plots is not an artefact of dimensionality reduction but reflects genuine structure in the underlying embedding space as this has been generated by the outgoing links of biographical pages.

Here it is worth talking about the clustering of centuries in broad historical periods since it has a specific implication. Historical periodisation (i.e., the division of the past into named eras) is a scholarly and cultural construct that serves mostly organisation

purposes and it is not a feature of history itself [45, 46]. Historians and philosophers of history have long debated the criteria by which periods are drawn, and the extent to which divisions like "the Middle Ages" or "the Renaissance" reflect genuine discontinuities or simply inherited conventions [47]. The fact that structurally similar groupings re-emerge here from the relational patterns of biographical cultural entities of scientists and philosophers, without being built into the method, suggests that these conventional divisions are not entirely arbitrary within the network's representational logic. They correspond to the conceptual landscapes that biographical articles share, and the cultural objects (e.g., ideas, institutions, intellectual traditions) that mediate their connections and that these conceptual landscapes are themselves distributed across time in ways that align with conventional periodisation.

This finding is important for two reasons. First, methodologically, it offers evidence that vector-space representations of large biographical networks preserve historically meaningful relational structure, which supports their use as tools for exploring how cultural knowledge is organised and transmitted at scale [48–50]. Second, at the interpretation level, it raises a question that the method alone cannot resolve: do these clusters emerge because the cultural entities embedded in Wikipedia's network share conceptual and relational contexts that following historical periodisation, or because the network has been constructed within a historiographical tradition that already organises knowledge in those terms? As mentioned in the *Introduction*, Wikipedia's cultural entities are not direct representations of historical reality, but abstract knowledge entities shaped by the editorial and epistemic regime through which they were admitted into the network (see *Introduction*).

The two explanations are not mutually exclusive, and the finding is consistent with both simultaneously. Intellectual history may have genuinely unfolded in periodised ways; editors working within a historiographical tradition may have organised their writing accordingly; or, most likely, both processes have reinforced one another over time. As argued in the *Introduction*, Wikipedia's cultural entities are abstract knowledge

entities shaped by the epistemic and editorial regime through which they were admitted into the network; a regime that is itself a product of the intellectual history it encodes. The clustering result therefore does not resolve the question of which explanation is primary. What it does show is that the relational structure of the network is consistent with conventional periodisation at the level of entire centuries, and that this consistency is recoverable through computational methods that impose no historical categories of their own.

Pairwise Semantic Similarity and the Temporal Tobler Pattern

While the centroid-based analysis offered a macroscopic view of how semantic profiles relate to historical time, it does so by aggregating information across many biographies simultaneously. This aggregation, by design, conceals variation at the level of the individual. To recover that finer-grained picture, the analysis now shifts to the level of individual nodes, in line with the principle of focus reading (see *Introduction*). However, this analysis departs slightly from the specific application of focus reading used in *section 3.1.2.3*, where individual nodes were examined to characterise the cultural network of Art, Science, and Philosophy in the 17th century. Here, the unit of analysis is not a single biography in isolation, but rather a pair of century-level centroid embeddings. Each centroid is constructed by pooling the outgoing Wikipedia links of all biographical pages whose subjects were born in a given century. Comparing the cosine similarity between century centroids against their temporal distance allows structural patterns to emerge that would remain invisible at the aggregate level. The pairwise analysis therefore makes it possible to ask a more precise version of the central question: not whether semantic profiles shift over historical time on average, but whether two centuries that are temporally distant also tend to be semantically distant, and whether this relationship holds consistently across the dataset.

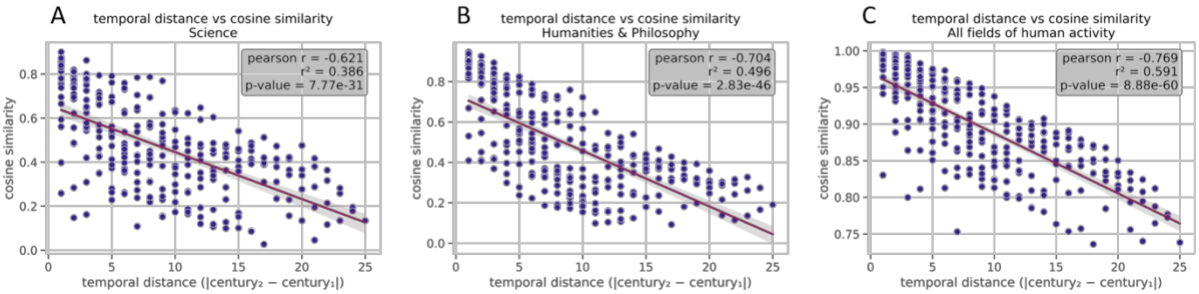


Figure 30 Relationship between temporal distance and semantic similarity for Science (**A**), Humanities & Philosophy (**B**), and all FHAs combined (**C**). The scatterplot shows the correlation between temporal distance (x-axis) and the cosine similarity of the embeddings of century centroids of biographies (y-axis).

The analysis shows that across the combined set of biographies and within individual FHAs (see **Figure 30**), cosine similarity between century centroids tends to decrease gradually with increasing temporal distance. In other words, centuries whose biographical corpora are closer together in time tend to reference more similar semantic contexts than those separated by many centuries. The strength of this relationship is, however, modest. The coefficients of determination are moderate (ranging from 0.39 to 0.59), indicating that temporal distance accounts for a substantial but partial share of the variation in semantic similarity, with considerable scatter remaining at the level of individual century pairs. By analogy with Tobler's first law of geography — the principle that near things tend to be more related than distant things [51]— this empirical regularity is referred to here as a "temporal Tobler" pattern: the tendency for temporal proximity to be associated with greater semantic proximity in the embedding space. As with its geographical counterpart, the pattern is not absolute; many century pairs that are temporally distant are semantically similar, and vice versa. What the pattern captures is a statistical tendency at the aggregate level, not a deterministic relationship at the level of individual pairs. The effect is moderate and consistent across FHAs, but some unexplained variance remains. Because that residual variation is not uniform, some century pairs deviate from the fitted trend, and these residual pairs are of particular interest, since the topics shared between them may point to concepts, individuals, or institutions whose prominence recurs across distant

periods. To surface these cases systematically, an outlier detection procedure is applied to the full set of pairwise century comparisons for the FHAs under study.

For each pair of century centroids, the cosine similarity between the two pooled embeddings is computed alongside their temporal distance. A linear trend is then fitted to the relationship between distance and similarity, and the residual of each pair is standardised into a z-score. Pairs are flagged as outliers if they are separated by at least 10 centuries (this is the average temporal distance in the dataset), and their residual z-score falls in the top five percent of all distant pairs. This residual-based criterion ensures that what is identified is not merely high similarity in absolute terms, but similarity that is anomalously high given the temporal distance involved.

For each outlier pair, the shared topics are retrieved by intersecting the outgoing link vocabularies of the two century corpora, which in this case is the full set of Wikipedia articles linked to from the biographical pages belonging to each century. These linked articles include not only other biographical pages but also articles about concepts, texts, institutions, and movements. Shared topics are ranked by their relative frequency across the two corpora rather than by raw co-occurrence counts. This normalisation corrects for differences in corpus size across centuries, which would otherwise inflate raw term frequencies in larger corpora independently of any historical signal. In other words, by working with proportions, the ranking reflects terms that are meaningfully prominent in both corpora, not merely abundant in the larger one. The shared topic vocabularies for each flagged pair are visualised as word clouds, in which font size is proportional to the minimum relative frequency of each term across the two corpora.

The procedure flags 8 outlier pairs in total: 2 for Science and 6 for Humanities & Philosophy. The two FHAs do not share the same endpoints which indicates that different FHAs encode distinct historical processes in their semantic profiles.

For Science, two pairs are flagged. The first is (-2, 13), connecting the 2nd century BCE to the 13th century CE across a temporal distance of 15 centuries, with a cosine

similarity of 0.905 (see **Figure 31A**). The second is (1, 13), connecting the 1st century CE to the 13th century CE across a temporal distance of 12 centuries with a cosine similarity 0.929 (see **Figure 31B**). Both pairs share the 13th century as their later endpoint, though the earlier endpoints differ.

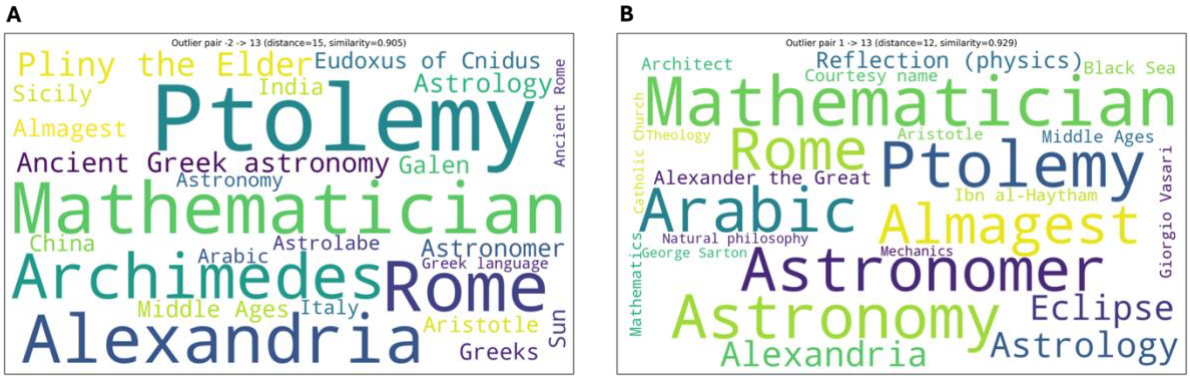


Figure 31 Word clouds showing the shared topic vocabulary for the two Science outlier pairs. **(A)** Outlier pair −2-13 (distance = 15, similarity = 0.905). **(B)** Outlier pair 1-13 (distance = 12, similarity = 0.929). In both panels, font size is proportional to the minimum relative frequency of each term across the two century corpora.

The shared vocabulary of the (-2, 13) pair is organised around a cluster of scientific figures of the Classical Era. The most visually prominent terms include Ptolemy, Mathematician, Archimedes, and Alexandria, with secondary terms including Astronomy, Astrology, Almagest, Arabic, Galen, Eudoxus of Cnidus, and Pliny the Elder. The presence of Archimedes and Eudoxus of Cnidus alongside Ptolemy locates the 2nd-century BCE endpoint firmly in the tradition of Hellenistic mathematical and astronomical science centred on Alexandria. The appearance of Arabic and Almagest alongside Middle Ages points toward the 13th-century endpoint, where this classical heritage was received and institutionalised through the Arabic-Latin translation movement [37, 52]. Galen and Pliny the Elder extend the vocabulary beyond astronomy into natural history and medicine, suggesting that the semantic bridge between these two centuries is not limited to a single scientific domain but spans the broader inheritance of Hellenistic natural inquiry.

The shared vocabulary of the (1, 13) pair is similarly organised around mathematical and astronomical transmission but with slightly different internal structure. The most prominent terms include Mathematician, Astronomy, Astronomer, Arabic, Almagest, and Ptolemy, with Ibn al-Haytham, Middle Ages, and Natural philosophy also present. Compared to the (−2, 13) pair, Archimedes and the specifically Hellenistic figures are less prominent, while Arabic and Ibn al-Haytham are more so. This shift is consistent with the 1st-century CE endpoint being less directly associated with the Hellenistic school of Alexandria and more with the broader Greco-Roman scientific tradition that the Arabic translators later engaged. The presence of Ibn al-Haytham alongside Ptolemy is consistent with the view that the Arabic tradition contributed actively to this transmission rather than serving only as a conduit [53, 54]. The appearance of Giorgio Vasari is most plausibly explained by the fact that biographical pages of Renaissance polymaths, which Vasari documented in his book *Le Vite de' più eccellenti pittori, scultori, e architettori* (*Lives of the Most Excellent Painters, Sculptors, and Architects)*, carry outgoing links that overlap with the classical and Arabic scientific vocabulary shared between the two century corpora.

Together, the two Science pairs point toward a historically recognisable process: the transmission and reception of classical scientific knowledge, organised around the mathematical and astronomical inheritance of the Hellenistic world, with the 13th century as its semantic concentration point in the medieval Latin tradition.

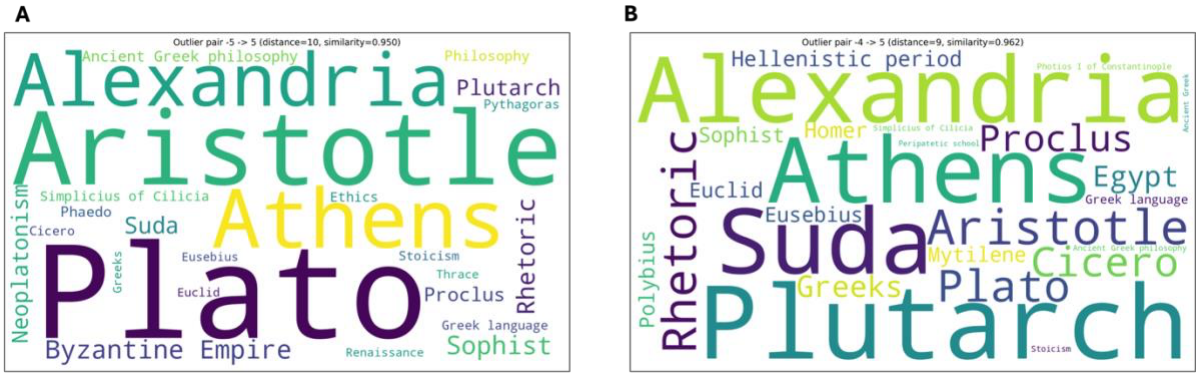


Figure 32 Word clouds showing the shared topic vocabulary for two Humanities & Philosophy outlier pairs. **(A)** Outlier pair −5-5 (distance = 10, similarity = 0.950). **(B)** Outlier pair −4-5 (distance = 9,

similarity = 0.962). In both panels, font size is proportional to the minimum relative frequency of each term across the two century corpora.

For Humanities & Philosophy, six pairs are flagged. Their endpoints and similarities are as follows: (−5, 5) at distance 10 and similarity 0.950 (see **Figure 32A**); (−4, 5) at distance 9 and similarity 0.962 (see **Figure 32B**); (1, 15) at distance 14 and similarity 0.915 (see **Figure 33A**); (4, 13) at distance 9 and similarity 0.947 (**Figure 33B**); (4, 14) at distance 10 and similarity 0.939 (see **Figure 34A**); and (4, 15) at distance 11 and similarity 0.939 (see **Figure 34B**).

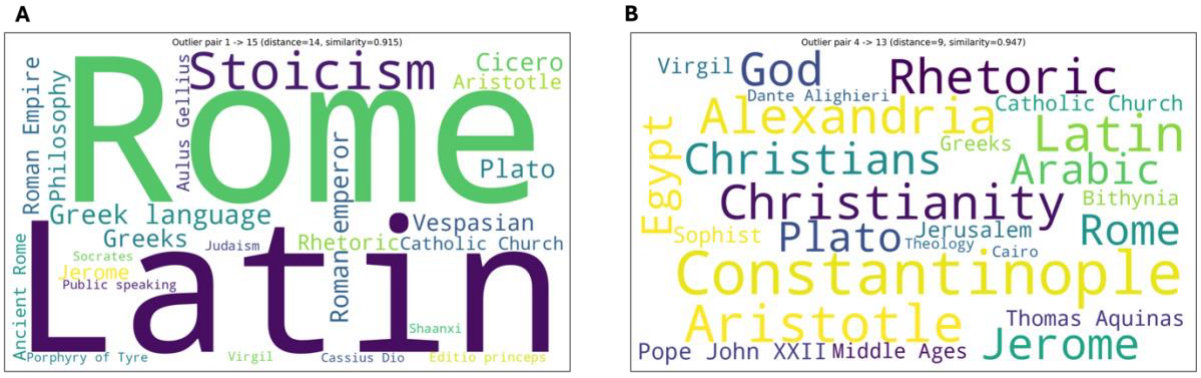


Figure 33 Word clouds showing the shared topic vocabulary for two Humanities & Philosophy outlier pairs. **(A)** Outlier pair 1-15 (distance = 14, similarity = 0.915). **(B)** Outlier pair 4-13 (distance = 9, similarity = 0.947). In both panels, font size is proportional to the minimum relative frequency of each term across the two century corpora.

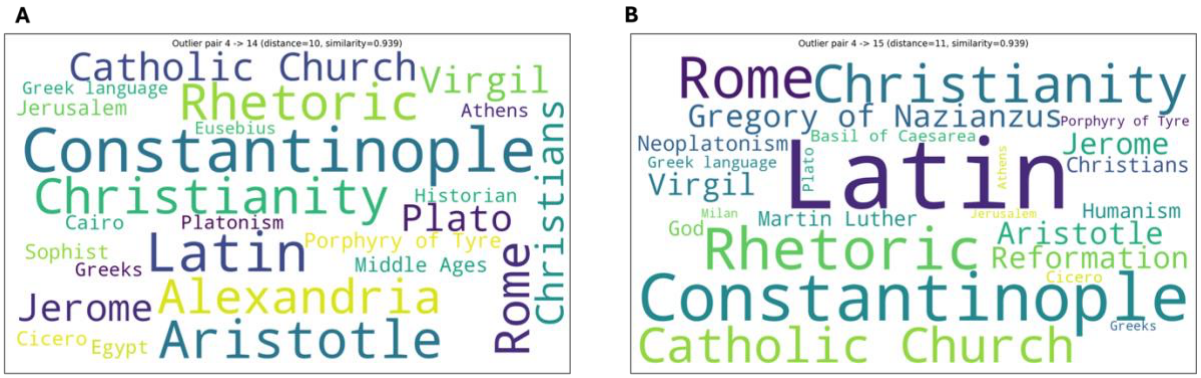


Figure 34 Word clouds showing the shared topic vocabulary for two Humanities & Philosophy outlier pairs. **(A)** Outlier pair 4-14 (distance = 10, similarity = 0.939). **(B)** Outlier pair 4-15 (distance = 11, similarity = 0.939). In both panels, font size is proportional to the minimum relative frequency of each term across the two century corpora.

The first group comprises the pairs (−5, 5) and (−4, 5), both of which connect a century from the Classical Greek period to the 5th century CE. The earlier endpoint of (−5, 5)

is the 5th century BCE, the period of Socrates and the Sophists; the earlier endpoint of (−4, 5) is the 4th century BCE, the period of Plato and Aristotle. The later endpoint in both cases is the 5th century CE, which belongs to Late Antiquity and is characterised by the systematic reception and commentary of the Platonic and Aristotelian corpus within Neoplatonist and early Christian frameworks [55].

The word cloud for (−5, 5) shows Alexandria and Aristotle as the most prominent terms, followed by Plato, Athens, Suda, Proclus, Neoplatonism, Simplicius of Cilicia, Byzantine Empire, Sophist, Rhetoric, and Ancient Greek philosophy. The word cloud for (−4, 5) shows a closely related vocabulary: Alexandria, Athens, Suda, Rhetoric, Plutarch, Proclus, Plato, Aristotle, Cicero, Simplicius of Cilicia, and Ancient Greek philosophy. Across both word clouds, the shared vocabulary is that of the late antique commentary tradition in which figures such as Proclus and Simplicius produced systematic commentaries on Plato and Aristotle through the schools of Athens and Alexandria, with encyclopaedic works such as the Suda serving as preserving and transmitting vehicles. The high cosine similarities of both pairs (0.950 and 0.962 respectively) are historically legible: the 5th century CE is semantically close to the Classical Greek centuries not because the two periods were contemporaneous, but because the intellectual project of the later period was organised substantially around the explication and preservation of the earlier one (see also **Figure 28A** where the 5th century CE appears very close to the Classical Era cluster).

The second group comprises the four pairs that involve the 4th century CE as their earlier endpoint: (4, 13), (4, 14), and (4, 15), as well as the pair (1, 15), which has the 1st century CE as its earlier endpoint and the 15th century CE as its later one.

The word cloud for (1, 15) (see **Figure 33A**) is dominated by Rome, Latin, and Stoicism, with Rhetoric, Greek language, Cicero, Plato, Aristotle, Jerome, and Catholic Church also present. The 1st-century CE endpoint corresponds to the Latin intellectual tradition and more specifically to the period of Cicero, Stoic philosophy, and the early

consolidation of Latin as a literary and scholarly language. The 15th-century CE endpoint belongs to the Renaissance, which engaged directly with this Latin classical inheritance alongside its Greek antecedents. The shared vocabulary of this pair thus maps onto the Roman strand of the classical tradition, as distinct from the more purely Greek vocabulary visible in the (−5, 5) and (−4, 5) pairs.

The word cloud for (4, 13) (see **Figure 33B**) is dominated by Constantinople, Aristotle, Alexandria, Christianity, Christians, Jerome, and Rhetoric, with Latin, Arabic, Plato, God, Catholic Church, and Thomas Aquinas also present. The 4th-century CE endpoint corresponds to Late Antiquity at the moment of Christianity's institutional consolidation and more specifically the period of the Council of Nicaea, Jerome's Vulgate translation, and the absorption of Platonic and Aristotelian philosophy into Christian theology [56, 57]. The 13th-century CE endpoint corresponds to the scholastic synthesis, when figures such as Thomas Aquinas engaged systematically with Aristotle through Arabic intermediaries.

The word cloud for (4, 14) (see **Figure 34A**) shows a vocabulary closely related to (4, 13) but with some differences in emphasis. Constantinople, Rhetoric, Christianity, Latin, Jerome, Alexandria, Aristotle, Plato, and Catholic Church are again prominent, but Porphyry of Tyre, Platonism, Sophist, Historian, and Cicero also appear, and the Christian-institutional terms are somewhat more balanced against the philosophical ones.

The word cloud for (4, 15) (see **Figure 34B**) is dominated by Constantinople, Catholic Church, Christianity, Latin, Rhetoric, Rome, and Jerome, with Gregory of Nazianzus, Basil of Caesarea, Neoplatonism, Virgil, Plato, Aristotle, Martin Luther, and Humanism also present. The vocabulary here is the most overtly Christian-institutional of the four pairs in this group, while also including humanist terms (Humanism, Virgil, Cicero) consistent with the 15th-century endpoint. The presence of Martin Luther is most plausibly explained by the birth-century assignment used throughout this analysis:

figures whose intellectual activity is associated with the early 16th-century Reformation were born in the 15th century and therefore contribute Reformation-related semantic content to the 15th-century corpus. Therefore, this should be read as a consequence of the classification method rather than as evidence that the Reformation is a conceptually central feature of the connection between the 4th and 15th centuries.

To summarize, the six Humanities & Philosophy pairs point toward two related but distinct historical processes. The first, captured by the (−5, 5) and (−4, 5) pairs, is the late antique reception of Classical Greek philosophy and more specifically the commentary tradition that preserved and transmitted Plato and Aristotle through the schools of Athens and Alexandria. The second, captured by the (1, 15), (4, 13), (4, 14), and (4, 15) pairs, is the longer arc of classical reception through Christian institutional and humanist frameworks, in which the 4th century CE serves as an early consolidation point and the 13th through 15th centuries CE as successive moments of re-engagement.

Together, the outliers in both FHAs almost exclusively correspond to historically recognised eras of intensified transmission and reception: the Islamic translation movement and its Latin scholastic uptake in the case of Science, and the late antique commentary tradition, early Christian institutionalisation, and humanist re-engagement in the case of Humanities & Philosophy. In this sense, the outliers can be read as empirical markers of what might be called temporal bridges. In other words, specific centuries that function as hubs of long-range knowledge transmission, where earlier corpora were received, reorganised, and reintroduced into new intellectual contexts. The fact that the vocabulary of these temporal bridges differs systematically across fields could be explained by the fact it reflects distinct regimes of transmission, in which Science is anchored by identifiable technical texts and their translation chains, while Humanities & Philosophy is mediated by a wider network of commentary traditions, institutional theology, and literary culture. However, it is worth noting, that a broader and more culturally mediated vocabulary is also more susceptible to

retrospective editorial framing: Wikipedia biographies of Late Antique or Renaissance figures tend to link outward to articles that reflect how modern scholarship categorises their historical significance, rather than concepts strictly contemporary to those figures. Whether the breadth of the Humanities & Philosophy vocabulary reflects the genuine historical complexity of classical transmission in that field, or partly an artefact of editorial framing, is a distinction the data alone cannot settle, and both interpretations should be held open.

3.2.3. Discussion

Summary of the findings. The analysis presented here converges on a core empirical claim: Wikipedia's biographical network encodes historical time in a structured and non-random way; a claim that the previous analysis already demonstrated at the scale of one century (see *Results – section 3.1*). More importantly, this structure is recoverable through both network and semantic methods applied at scale. Three findings support this claim, and they are presented below before their implications are drawn out.

First, temporal homophily operates as a general organising principle across the entire biographical network. Quantitatively, 96.99% of all biographical links connect individuals born within a one-century span, and the average temporal distance between linked individuals (i.e., 33.78 years) closely approximates the demographic interval of a human generation (27 to 30 years) [21, 22]. This pattern also holds within disciplinary subsets such as Science and Humanities & Philosophy. This result is consistent with the findings of Jatowt et al. [7], who identified temporal homophily over a narrower time window, but the present analysis extends it across 25 centuries and nearly two million biographical pages. This extension demonstrates that the pattern is not period specific. The convergence with the generational interval suggests that Wikipedia's link structure registers the temporal radius of social co-existence, which can be shaped by kinship, succession, collaboration, rivalry, etc., as refracted through encyclopaedic documentation.

Second, the minority of long-range connections is not randomly distributed and carries meaningful information. The directional asymmetry between incoming and outgoing based normalisation reveals what is described here as a topological bypass under outgoing normalisation, Modern Era scientists link preferentially to Classical Era figures, bypassing medieval intermediaries. Under incoming normalisation, however, the Middle Ages appear as a genuine connector, receiving links from the Classical Era and transmitting them forward in time. This finding is consistent with the broader argument in digital humanities that quantitative representation of cultural data encode interpretive choices that shape what become visible [44].

Third, the embeddings analysis shows that temporal organization extends beyond the links structure into the semantic content of biographical pages. Century-level centroid vectors form partially overlapping clusters that correspond to conventional historical periods, without those periods being built into the method. This correspondence is confirmed by hierarchical clustering, cluster quality metrics (Silhouette scores of 0.33 for Science, 0.26 for Humanities & Philosophy and 0.22 for all FHAs combined), and permutation tests in which none of 10,000 random rearrangements of period labels produced an equal or higher score ($p = 0.0001$). At the focus reading level and more specifically at the pairwise comparison, cosine similarity decreases gradually with increasing temporal distance across all FHAs examined, a regularity described here as a temporal Tobler pattern by analogy with Tobler's first law of geography [51]. Lastly, the outlier detection procedure applied to the pairwise comparisons identifies century pairs whose semantic similarity is anomalously high relative to their temporal distance. The procedure flags four such pairs in total, and their distribution is itself analytically significant: the Science outliers cluster around the 13th century CE with a vocabulary related to the Islamic Golden Age, while the Humanities & Philosophy outlier clusters around the 15th century CE with a vocabulary related to the Renaissance and the Christian clergy.

Theoretical Implications. The results presented above have a specific implication for how large-scale digital repositories can be used as tools for examining questions about cultural memory and disciplinary epistemology.

The first implication concerns the relationship between network topology and semantic content. The network analysis and the embeddings analysis address different levels of the same object; the link structure and the conceptual context of links, respectively. They converge on a consistent picture: temporal organisation is present at both levels, but in different forms. Network topology shows sharp concentration and selective long-range bypasses, whereas semantic content shows gradual, graded change. That pattern suggests that the cultural network of scientists and philosophers does not encode a single temporal logic but two simultaneously. Treating these two approaches as complementary rather than redundant, as the framework adopted here does, allows patterns to emerge that neither analysis could surface alone.

The second contribution engages with the historiographical content embedded in the network's structure, and in doing so returns to the concept of cultural entities introduced in the *Introduction* of this thesis. As established there, a cultural entity in the context of Wikipedia is not a direct representation of its real-world referent but a constructed knowledge object consisting of two analytically distinct layers. The first is a historiographical layer: the accumulated record produced by expert scholarship through the documentation and interpretation of past events and figures. The second is an editorial layer: the re-interpretation and selection of that expert record by Wikipedia's community of editors, operating under the platform's normative guidelines of verifiability, notability, and neutrality. These two layers are entangled in every article and, by extension, in every edge of the wikilink network. This entanglement has direct consequences for how the network patterns identified in this section should be interpreted. The topological bypass finding shows that the same network simultaneously supports a reading of historical rupture and a reading of continuity depending on the normalisation applied. In my opinion, this divergence is

not a technical artefact. The two normalisation procedures weight the network's edges differently, and in doing so they effectively "privilege" different layers of the cultural entity: one amplifies the traces of the historiographical record, the other amplifies the editorial decisions through which that record was selectively incorporated into Wikipedia. The case of the Middle Ages illustrates this concretely. The relative peripherality of medieval figures under outgoing normalisation does not mean that the Middle Ages are historiographically suppressed in Wikipedia. Rather, the connections through which medieval knowledge is integrated into the network are foregrounded differently depending on which normalisation is applied and, by extension, which layer of the cultural entity is being read. From one vantage point, the intellectual contributions of the Middle Ages appear marginal; from another, they appear as structurally embedded connective tissue linking ancient and early modern knowledge. This distinction matters not only for the empirical analysis presented here but for the theoretical treatment of digital cultural repositories more broadly: what a knowledge network reveals about the past is always a function of how the network is read and analysed, not only of what the network contains.

Practical Implications. That last theoretical implication is the connector to a practical implication of this study. More specifically, the methodological finding that normalisation direction shapes the historical narrative made legible by the data has a direct consequence for researchers working with large amount of information and using models to abstract this information. As Ahnert et al. argued, any network is an abstraction thus the choices we make when building this abstraction affects the analysis itself [58]. In general, in many cultural studies there is always, sometimes implicitly, a choice of directionality (e.g., which periods influenced which, which figures are central etc.) The present analysis suggests that both normalisation perspectives should be applied and reported as a matter of standard practice. Moreover, this is not only important for Wikipedia studies but for every study using network data to quantify culture and measure some cultural phenomena.

Lastly, the last practical point that should be highlighted is the use of semantic embeddings in cultural analytics. The permutation-tested Silhouette scores reported here provide a concrete model for validating whether period-level clustering in embedding space reflects genuine structure or dimensional compression artefacts. Applying this verification sequence (i.e., PCA projection, hierarchical clustering, cluster quality metrics, and permutation testing) guards against the over-interpretation of visual patterns in reduced-dimensional projections and should be considered in studies of digital humanities and cultural analytics.

Bibliography

1. Eom Y-H, Shepelyansky DL (2013) Highlighting entanglement of cultures via ranking of multilingual Wikipedia articles. PLOS ONE 8:1–10. https://doi.org/10.1371/journal.pone.0074554

2. Eom Y-H, Aragón P, Laniado D, et al (2015) Interactions of Cultures and Top People of Wikipedia from Ranking of 24 Language Editions. PLOS ONE 10:e0114825. https://doi.org/10.1371/journal.pone.0114825

3. Aragón P, David Laniado, Laniado D, et al (2012) Biographical social networks on Wikipedia: a cross-cultural study of links that made history. WikiSym '12 19. https://doi.org/10.1145/2462932.2462958

4. Frahm KM, Jaffrès-Runser K, Shepelyansky DL (2016) Wikipedia mining of hidden links between political leaders. arXiv: Social and Information Networks. https://doi.org/10.1140/epjb/e2016-70526-3

5. Chen B, Lin Z, Evans TS (2019) Analysis of the Wikipedia Network of Mathematicians

6. Rollin G, Lages J (2025) The most influential philosophers in Wikipedia: a multicultural analysis. EPJ Data Sci 14:69. https://doi.org/10.1140/epjds/s13688-025-00570-w

7. Jatowt A, Kawai D, Tanaka K (2016) Digital History Meets Wikipedia: Analyzing Historical Persons in Wikipedia. In: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries. Association for Computing Machinery, New York, NY, USA, pp 17–26

8. Reznik I, Shatalov V (2016) Hidden revolution of human priorities: An analysis of biographical data from Wikipedia. Journal of Informetrics 10:124–131. https://doi.org/10.1016/j.joi.2015.12.002

9. Miccio LA, Gámez-Pérez C, Suárez JL, Schwartz GA (2022) Mapping the Networked Context of Copernicus, Michelangelo, and Della Mirandola in Wikipedia. Advs Complex Syst 25:2240010. https://doi.org/10.1142/S0219525922400100

10. Miccio LA, Agapitos P, Gamez-Perez C, et al (2025) Wikipedia as a cultural lens: a quantitative approach for exploring cultural networks. Humanit Soc Sci Commun 12:462. https://doi.org/10.1057/s41599-025-04772-5

11. Schwartz GA (2021) Complex networks reveal emergent interdisciplinary knowledge in Wikipedia. Humanit Soc Sci Commun 8:1–6. https://doi.org/10.1057/s41599-021-00801-1

12. Sangiacomo A, Tanasescu R, Hogenbirk H, Donker S (2022) Recreating the Network of Early Modern Natural Philosophy: A Mono- and Multilingual Text Data Vectorization Method. Journal of Historical Network Research 33-85 Pages. https://doi.org/10.25517/JHNR.V7I1.129

13. Stoltz DS, Taylor MA (2021) Cultural cartography with word embeddings. Poetics 88:101567. https://doi.org/10.1016/j.poetic.2021.101567

14. Lee M, Martin JL (2015) Coding, counting and cultural cartography. Am J Cult Sociol 3:1–33. https://doi.org/10.1057/ajcs.2014.13

15. Kozlowski AC, Taddy M, Evans JA (2019) The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. Am Sociol Rev 84:905–949. https://doi.org/10.1177/0003122419877135

16. Garg N, Schiebinger L, Jurafsky D, Zou J (2018) Word embeddings quantify 100 years of gender and ethnic stereotypes. Proc Natl Acad Sci USA 115:. https://doi.org/10.1073/pnas.1720347115

17. Deb S, Chanda AK (2022) Comparative analysis of contextual and context-free embeddings in disaster prediction from Twitter data. Machine Learning with Applications 7:100253. https://doi.org/10.1016/j.mlwa.2022.100253

18. Ethayarajh K (2019) How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp 55–65

19. Liu Y, Tilahun G, Gao X, et al (2024) Comparative Analysis of Static and Contextual Embeddings for Analyzing Semantic Changes in Medieval Latin Charters

20. VanderPlas J (2017) Python data science handbook: essential tools for working with data, First edition. O'Reilly, Beijing Boston Farnham Sebastopol Tokyo

21. Tremblay M, Vézina H (2000) New Estimates of Intergenerational Time Intervals for the Calculation of Age and Origins of Mutations. The American Journal of Human Genetics 66:651–658. https://doi.org/10.1086/302770

22. Wang RJ, Al-Saffar SI, Rogers J, Hahn MW (2023) Human generation times across the past 250,000 years. Sci Adv 9:eabm7047. https://doi.org/10.1126/sciadv.abm7047

23. McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a Feather: Homophily in Social Networks. Annu Rev Sociol 27:415–444. https://doi.org/10.1146/annurev.soc.27.1.415

24. Glänzel W, Schoepflin U (1995) A bibliometric study on ageing and reception processes of scientific literature. Journal of Information Science 21:37–53. https://doi.org/10.1177/016555159502100104

25. Larivière V, Archambault É, Gingras Y (2008) Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004). J Am Soc Inf Sci 59:288–296. https://doi.org/10.1002/asi.20744

26. Chang Y-W (2013) A comparison of citation contexts between natural sciences and social sciences and humanities. Scientometrics 96:535–553. https://doi.org/10.1007/s11192-013-0956-1

27. Gilyarevskii RS, Libkind AN, Libkind IA, Bogorov VG (2021) The Obsolescence of Cited and Citing Journals: Half-Lives and Their Connection to Other Bibliometric Indicators. Autom Doc Math Linguist 55:152–165. https://doi.org/10.3103/S0005105521040026

28. Suarez J-L, McArthur B, Soto-Corominas A (2015) Cultural Networks and the Future of Cultural Analytics. In: 2015 International Conference on Culture and Computing (Culture Computing). IEEE, Kyoto, Japan, pp 95–98

29. Sorokin PA, Merton RK (1937) Social Time: A Methodological and Functional Analysis. American Journal of Sociology 42:615–629. https://doi.org/10.1086/217540

30. Adam B (1994) Time and social theory, 1. publ. in pbk. Polity Press, Cambridge

31. Saraceno C (1989) The time structure of biographies. enquete. https://doi.org/10.4000/enquete.80

32. Brennecke J, Carnabuci G, Ertug G (2025) Relational History: Correcting temporal myopia in social network research. Organization Theory 6:26317877251379502. https://doi.org/10.1177/26317877251379502

33. Moretti F (2000) Conjectures on World Literature. NLR 2:54–68. https://doi.org/10.64590/hxj

34. Le Goff J (2015) Must we divide history into periods? Columbia University Press, New York Chichester

35. Grant E (1996) The Foundations of Modern Science in the Middle Ages: Their Religious, Institutional and Intellectual Contexts, 1st ed. Cambridge University Press

36. Hannam J (2010) God's Philosopher: How the Medieval World Laid the Foundations of Modern Science. Faber and Faber, London

37. Montgomery SL (2018) Mobilities of Science: The Era of Translation into Arabic. Isis 109:313–319. https://doi.org/10.1086/698236

38. Ragab A (2022) Translation and the Making of a Medical Archive: The Case of the Islamic Translation Movement. Osiris 37:25–46. https://doi.org/10.1086/719219

39. Brentjes S, Morrison R (2010) The sciences in Islamic societies (750–1800). In: Irwin R (ed) The New Cambridge History of Islam: Volume 4: Islamic Cultures and Societies to the End of the Eighteenth Century. Cambridge University Press, Cambridge, pp 564–639

40. Saatsaz M, Rezaei A (2023) The technology, management, and culture of water in ancient Iran from prehistoric times to the Islamic Golden Age. Humanit Soc Sci Commun 10:142. https://doi.org/10.1057/s41599-023-01617-x

41. Needham P (2020) Getting to know the world scientifically: an objective view. Springer, Cham

42. Pentzold C (2009) Fixing the floating gap: The online encyclopaedia Wikipedia as a global memory place. Memory Studies 2:255–272. https://doi.org/10.1177/1750698008102055

43. Wagner C, Garcia D, Jadidi M, Strohmaier M (2015) It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. Proceedings of the International AAAI Conference on Web and Social Media 9:454–463. https://doi.org/10.1609/icwsm.v9i1.14628

44. Drucker J (2011) Humanities Approaches to Graphical Display. Digital Humanities Quarterly 5:. https://doi.org/10.63744/r4ysrh7ae534

45. Shaw R (2020) Periodization. https://www.isko.org/cyclo/periodization. Accessed 31 May 2026

46. Sadeghi Y, Islam G, Van Lent W (2025) Practices of Periodization: Toward a Critical Perspective on Temporal Division in Organizations. AMR 50:51–71. https://doi.org/10.5465/amr.2022.0396

47. Guldi J, Armitage D (2014) The History Manifesto, 1st ed. Cambridge University Press

48. Hamilton WL, Leskovec J, Jurafsky D (2016) Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change

49. Kutuzov A, Øvrelid L, Szymanski T, Velldal E (2018) Diachronic word embeddings and semantic shifts: a survey. In: Bender EM, Derczynski L, Isabelle P (eds) Proceedings of the 27th International Conference on Computational Linguistics. Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp 1384–1397

50. Martinc M, Kralj Novak P, Pollak S (2020) Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift. In: Calzolari N, Béchet F, Blache P, et al (eds) Proceedings of the Twelfth Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp 4811–4819

51. Tobler WR (1970) A Computer Movie Simulating Urban Growth in the Detroit Region. Economic Geography 46:234. https://doi.org/10.2307/143141

52. El Ghazi O, Bnini C (2020) Arabic Translation from Bait Al-Hikma to Toledo School of Translators: Key Players, Theorization and Major Strategies. IJLLT 3:66–80. https://doi.org/10.32996/ijllt.2020.3.9.7

53. Abudawood GA, Alshareef R, Alghamdi S (2020) Arab and Muslim scientists and their contributions to the history of ophthalmology. Saudi Journal of Ophthalmology 34:198–201. https://doi.org/10.4103/1319-4534.310407

54. Sedgwick HA (2022) Ibn al-Haytham's ground theory of distance perception. i-Perception 13:20416695221118388. https://doi.org/10.1177/20416695221118388

55. Chistyakova O, Chistyakov D (2023) Theology of Greek-Byzantine Church Fathers as a Specific Way of Philosophical Thinking in an Epistemological Context. Religions 14:355. https://doi.org/10.3390/rel14030355

56. Roelli P (2021) Latin as the Language of Science and Learning. De Gruyter, pp 290–308

57. Miller DG (2012) Continuity and revival of classical learning. In: External Influences on English, 1st ed. Oxford University PressOxford, pp 192–227

58. Ahnert R, Ahnert SE, Coleman CN, Weingart SB (2020) 5 Quantifying Culture. In: The Network Turn: Changing Perspectives in the Humanities, 1st ed. Cambridge University Press, pp 73–88

4. Conclusions, Limitations & Future Work

4.1. Overview

The goal of this thesis was to develop and apply methods, both newly created and already existing, to identify, quantify, and interpret connectivity patterns between cultural entities, using as a case study the network of wikilinks of the English Wikipedia. That aim was pursued through three successive contributions, each building on the previous one. The first contribution examined whether an analysis of the network of Wikipedia can recover culturally meaningful connectivity patterns between cultural entities within a defined historical period and specific disciplines. As a proof of concept, I focused on the network of Art, Science, and Philosophy in the 17th century Europe (see *Results – section 3.1*). The second contribution addressed the computational infrastructure required to extend such analysis beyond a single historical period, presenting WikiTextGraph, an open-source tool for extracting text and the wikilinks network from Wikipedia dumps in a reproducible and efficient manner (see *Methodology – section 2.3*). The third brought together the analytical insights of the first contribution and the software developed in the second to propose a large-scale framework combining network and distributional semantics approaches, through which temporal and semantic patterns across approximately 2 million biographies spanning 25 centuries were identified and interpreted.

The present section draws together the main findings of this thesis and presents their collective theoretical and practical implications. It also reflects on the limitations of the work and qualifies the scope of these findings. Throughout, the section maintains the interpretive frame established in the *Introduction*: the network of Wikipedia is not a neutral map of cultural history, but a representation shaped by two distinct layers of mediation. The first is historiographical: the patterns it encodes reflect centuries of

expert scholarly judgment about which figures, events, and ideas are historically significant. The second is editorial: Wikipedia's collective of editors has made further decisions about which of that knowledge merits inclusion in the platform and how it should be connected. The patterns identified in this thesis should therefore be read as evidence of how cultural knowledge has been doubly filtered and selectively encoded in a large, collaboratively constructed digital knowledge system, and not as direct evidence of historical reality.

4.2. Conclusions

4.2.1. Summary of Findings by Research Question

4.2.1.1. RQ1: Network Analysis of a Historical Period

The first research question asked whether a network-based analysis of Wikipedia's network structure can recover meaningful connectivity patterns between cultural entities, using a given historical period as a case study. The analysis of the 17th-century cultural network of Art, Science, and Philosophy answers this question affirmatively, demonstrating that Wikipedia's link structure does retain historically and culturally meaningful signals (see *Results - Wikipedia as a Cultural Lens – a Network Approach*).

The contribution that distinguishes this work from prior studies in the field is the shift from analysing individual triads of figures to averaging network metrics across many sampled triads. Rather than characterising the 17th century through the lens of a single configuration of individuals, which would reflect one particular selection of figures, this approach derives a portrait of the period from the averaged metrics of multiple triads, each representing a different combination of actors within Art, Science, and Philosophy. This makes the resulting characterisation of the historical period more robust and less contingent on any individual selection, addressing a recurring limitation in network-based cultural history research.

Four findings emerged from this analysis. First, the degree to which the 17th-century network separates into three disciplinary clusters sits between the values recovered for the Italian Renaissance and the early 20th century, consistent with the hypothesis that disciplinary specialisation has increased progressively across the modern period. Second, at the cluster level, Philosophy emerges as the most internally cohesive discipline and the Science–Philosophy connection as the strongest cross-disciplinary bond in the network, reflecting the well-documented reorientation of early modern Philosophy towards epistemology and the foundations of scientific knowledge. Third, Art is by far the most inward-looking discipline, with openness values that vary systematically with the geographical positioning of individual artists, consistent with the documented regional fragmentation of the 17th-century art world. Fourth, at the node level, within-discipline and cross-discipline connectivity follow different statistical distributions, suggesting that the two types of connection are generated by different underlying mechanisms.

In summary, these findings demonstrate that Wikipedia's network is sensitive to historically meaningful structural change and that the triad-averaging method can surface culturally central figures and concepts that were not selected at the outset. This last capacity aligns directly with the analytical logic of focus reading: large-scale structural analysis identifies what is worth examining closely before the close examination begins.

4.2.1.2. RQ2: A Scalable Pipeline for Wikipedia Data Extraction

The second research question asked how a scalable, multilingual pipeline for extracting the text and the network from Wikipedia's content articles can be designed and implemented in a way that remains accessible to non-technical audiences. To address this question, it is necessary to first situate it within the trajectory of the broader project.

This research question did not arise independently; it emerged directly from the findings of the first contribution. Once the period-specific analysis demonstrated that Wikipedia's link network, when examined through the methods of complex network analysis, can recover historically meaningful structural patterns, the logical next step was to ask whether a comparable approach could be extended to longer temporal windows and larger populations of articles. Scaling up the analysis in this way, however, required infrastructure that the standard online request methods provided by Wikipedia could not reliably support. While querying Wikipedia's API is appropriate for retrieving a small number of articles, this approach becomes increasingly impractical as the number of content pages grows, due to rate limits and the cumulative overhead of repeated individual requests.

To resolve this constraint, WikiTextGraph was developed. Rather than relying on live online queries, WikiTextGraph parses and processes a complete Wikipedia database dump, extracting both the textual content of articles and their internal wikilink structure to construct a network representation at scale. This infrastructure does two things simultaneously: (1) it resolves a technical bottleneck, and (2) it establishes the analytical conditions under which the third contribution becomes possible.

4.2.1.3. RQ3: Temporal and Semantic Patterns in the Biographical Subnetwork

The third research question asked what temporal and semantic patterns emerge in the biographical subnetwork of the English Wikipedia when analysed across twenty-five centuries. To address it, both network and distributional semantics methods were applied to a large dataset of approximately 2 million biographical pages, each classified by Field of Human Activity (FHA). This question can be seen as the continuation of the two preceding contributions: the first established that Wikipedia's link structure encodes historically meaningful signals at the period level, and the second provided the computational infrastructure to operate at a scale that would have otherwise been technically unfeasible. The third contribution brings these two

strands together and asks whether the patterns recovered in a single, bounded historical period hold and how they evolve when the analysis is extended to a larger temporal range of Wikipedia's biographical coverage by combining network and semantic methods. Network analysis and distributional semantics address different dimensions of the same object: the former captures the relational structure between biographical entities, while the latter captures the thematic and conceptual content of the articles themselves. Using them in tandem allows patterns to emerge that neither approach could surface in isolation.

Three principal findings resulted from this combined approach. The first is that temporal homophily operates as a general organising principle of the biographical network: biographical wikilinks connect overwhelmingly to individuals from the same or adjacent centuries, with an average temporal distance that approximates the length of a human generation. This extends and generalises a pattern previously reported in the literature over narrower time windows, establishing it as a structural property of the network across twenty-five centuries rather than a period-specific feature.

The second finding concerns the long-range connections that sit outside this dominant pattern and what they reveal about how Wikipedia encodes intellectual history. Examining the network from two complementary normalisation perspectives shows that the same data simultaneously supports a reading of historical continuity and a reading of historical rupture. This divergence is not a technical artefact but a reflection of the layered nature of the cultural entity, in which the historiographical record and the editorial decisions that incorporated it into Wikipedia leave distinguishable traces depending on the analytical perspective applied (see *Introduction*).

The third finding is that temporal organisation extends beyond link structure into the semantic content of biographical pages. Without any historical categories being built into the method, the semantic profiles of biographical pages organise themselves into clusters that correspond closely to conventional historical periods. At the pairwise

level, semantic similarity decreases gradually with temporal distance, a regularity described here as a temporal Tobler pattern. The outlier detection procedure applied to the full set of pairwise century comparisons identifies specific century pairs whose semantic proximity is anomalously high relative to their temporal distance. These outliers are of interest because they flag cases where biographical corpora separated by ten or more centuries share a concentration of the same figures, texts, and concepts. This pattern is consistent with long-range processes of knowledge transmission that would not be visible in a period-by-period analysis. What makes this finding worth noting is that these connections are recovered from the link structure of biographical pages without any prior historical categorisation built into the method; the correspondence with processes that historians have independently documented therefore serves as a form of external validation. The results showed that there is a disciplinary contrast between the two FHAs: the Science outliers are organised around a relatively delimited vocabulary of textual and figural transmission, while the Humanities and Philosophy outliers span a broader network of philosophical, theological, rhetorical, and literary culture. This difference may reflect genuine asymmetries in how the two fields encoded and transmitted their classical inheritance, though the possibility that this breadth is partly an artefact of Wikipedia's retrospective editorial framing cannot be excluded.

4.2.2. Theoretical and Practical Contributions

This thesis contributes to the field of cultural analytics by developing and applying a quantitative framework that treats Wikipedia's hyperlink network as a cultural network. The contribution is twofold: methodological and empirical. At the methodological level, this thesis introduces a software pipeline for analysing Wikipedia data dumps and their relational structure. At the empirical level, two interconnected results chapters demonstrate that this network can be analysed systematically across different scales of resolution, from the characterisation of a single historical period to the longitudinal examination of nearly two million biographical pages spanning twenty-

five centuries. Together, these contributions support the central claim that large-scale digital repositories such as Wikipedia do not merely store cultural information; they organise it in ways that are amenable to formal analysis and that surface patterns inaccessible to traditional humanistic methods operating at smaller scales.

The first theoretical contribution concerns the analytical status of historical periods as units of analysis. In much of the existing literature, structural or cultural properties of a given period are inferred from the analysis of individual figures, rendering the resulting characterisation sensitive to the particular composition of the sample. In other words, a different selection of individuals could, in principle, yield a substantially different picture.

This thesis addresses that limitation through the triad-averaging framework. Rather than analysing any single artist, scientist, or philosopher, the method draws a large number of sampled triads of figures from within the same century, computes network metrics for each, and derives period-level estimates by averaging across all sampled configurations. This procedure shifts the unit of analysis from the individual to the period itself. Structural properties such as the degree of disciplinary specialisation, the relative internal cohesion of constituent disciplines, and the strength of cross-disciplinary connections are thereby treated as statistical regularities of the period rather than as properties of any particular configuration of individuals. The historical period, in this framework, becomes a measurable and comparable analytical object in its own right.

The results for the seventeenth-century network demonstrate that this approach recovers patterns consistent with established historical knowledge. These include the centrality of the Science-Philosophy axis and the progressive decoupling of Art from both disciplines following the Italian Renaissance. At the same time, the methodology produces genuinely new quantitative regularities, including the power-law distribution of inter-cluster connectivity and the core-periphery organisation of intra-cluster

strength. These findings would not be visible from an individual-level analysis, which supports the argument that the triad-averaging procedure is not merely a methodological convenience but a substantive analytical advance.

The second theoretical contribution concerns the relationship between two distinct levels at which the same cultural network can be analysed: its topology (that is, the structure of its connections) and its semantic content (that is, the meaning encoded in the text of its pages). The biographical network analysis shows that temporal organisation is present at both levels, but in different forms.

At the topological level, the dominant pattern is temporal homophily: figures tend to be connected to others who lived in the same period. However, this general tendency is interrupted by selective but historically meaningful connections across centuries and eras. One notable finding is that, under certain analytical perspectives, medieval figures become structurally invisible, a result that raises questions about whose contributions the network systematically marginalises. At the semantic level, the pattern is one of gradual, graded change. Temporally adjacent centuries tend to occupy closer positions in embedding space, and conventional historical periods emerge as distinct semantic clusters without being imposed on the method in advance.

Building on this structural picture, the outlier analysis adds a finer layer of interpretation. By identifying century pairs whose semantic similarity is unexpectedly high relative to their temporal distance, it points to specific historical processes as the drivers of unexpected semantic affinity across centuries. In the case of Science, the Islamic Translation movement is identified as one such driver; in the case of Humanities and Philosophy, it is the humanist recovery of classical textual culture.

The convergence of network and semantic analyses on a consistent, if not identical, picture of temporal organisation supports the argument that these two dimensions are not redundant. Rather, they are mutually illuminating. Treating them as complementary is methodologically more powerful than applying either in isolation,

because each dimension reveals aspects of cultural organisation that the other cannot capture alone.

The practical contributions of this thesis operate at two levels. For researchers working with Wikipedia or comparable knowledge repositories, the results show that the wikilinks network can function as a data source for cultural and historical analysis, rather than merely as a supplement to the textual content of Wikipedia's articles. The structural patterns recovered here align with well-established historical knowledge across multiple periods and disciplines. This alignment supports the practical use of wikilinks networks as a meaningful empirical object in their own right, not merely a convenient proxy for other sources of evidence.

This finding has implications beyond Wikipedia itself. As cultural datasets continue to grow in scale, encompassing digitised archives, biographical databases, citation networks, and large-scale encyclopaedic corpora, the question of how to extract historically meaningful signal from relational structures shared across millions of entities becomes increasingly central to the practice of cultural analytics. This thesis demonstrates that such extraction is possible, and that it can be done in a way that is both computationally tractable and interpretively accountable. By this the thesis means that the patterns recovered can be evaluated against independent historical knowledge rather than treated as self-validating outputs of algorithmic processing.

4.3. Limitations

A central limitation of both analyses concerns the status of Wikipedia as the main data source. The results demonstrate that Wikipedia’s cultural network encodes historically meaningful structural and semantic patterns; however, what is captured is not cultural history itself, but a socially situated representation of it. As noted in the *Introduction*, Wikipedia is produced by a self-selected community of editors operating under norms of verifiability, notability, and neutrality, and grounded in pre-existing institutionalised knowledge. The patterns recovered here must therefore be interpreted as properties

of a historically mediated knowledge system rather than of the past directly. The following discussion outlines the principal sources of bias that shape this representation.

One of the most extensively documented biases concerns gender. Female subjects remain under-represented in Wikipedia's biographical corpus, and existing articles tend to be shorter, less interconnected, and more peripherally positioned than those of comparable male figures [1–3]. This imbalance is historically uneven, becoming more pronounced prior to the twentieth century, where women's contributions to Art, Science, and Philosophy were systematically excluded from the records on which Wikipedia relies. In the 17th-century analysis, the triads used to characterise the cultural network are therefore drawn from a predominantly male pool. Consequently, the observed structural properties (e.g., modularity, assortativity, and node frequency) reflect a network already filtered by this exclusion. Whether a more gender-balanced corpus would yield different structural patterns remains an open question.

Geographic bias introduces a second layer of distortion. Wikipedia's coverage is disproportionately concentrated on Western Europe and, increasingly, North America, with comparatively limited representation of other regions [4, 5]. Accordingly, the 17th-century network analysed here is more precisely a representation of Western European intellectual activity as encoded in the English-language edition of Wikipedia. Even within Europe, peripheral regions and minority traditions are likely under-represented relative to established centres of scholarly attention. This skew extends to the large-scale biographical network, where temporal homophily and semantic clustering primarily reflect individuals embedded in Anglophone-accessible historiographical traditions.

This leads to a further constraint: the exclusive reliance on the English-language Wikipedia. While methodologically justified by its scale, this choice introduces edition-specific biases tied to the perspectives and priorities of its predominantly Anglophone

editor base [6, 7]. Figures more extensively documented in other language editions may appear in attenuated form or be absent altogether [8]. This is particularly relevant for the identification of cross-period semantic bridges, such as the Islamic Translation Movement linking the Classical Era to the 13th century. Although the results align with established historiography, multilingual analyses could plausibly modify the prominence or configuration of these connections. Prior work shows that different language editions produce systematically different prominence rankings, indicating that the observed patterns are partly contingent on the chosen edition [9].

Wikipedia's notability criteria further constrain the composition of the network. The requirement for coverage in reliable secondary sources ties inclusion closely to the canon established by academic and journalistic institutions [2, 10–12]. As a result, the network systematically privileges figures who have been recognised and documented within these traditions, while excluding those who were marginalised, insufficiently recorded, or subsumed under collective forms of attribution [6]. The resulting structure reflects not cultural production as such, but its retrospective canonisation.

A related constraint arises from the uneven survival and digitisation of historical sources. Coverage of earlier periods depends on the availability of secondary materials that have persisted and are accessible to editors [13]. This introduces a non-random survivorship bias favouring institutionally embedded individuals and regions with durable archival infrastructures. Consequently, the networks reconstructed for the Classical Era, Middle Ages, and Early Modern period are conditioned by long-term processes of preservation and loss that cannot be fully corrected.

Temporal asymmetries in Wikipedia's own development add a further layer of unevenness. Since its founding in 2001, coverage has expanded incrementally, with recent periods and prominent figures typically receiving earlier and denser attention. Article content reflects not only accumulated historical knowledge but also the evolving priorities and demographic composition of the editor community [14]. For an

analysis spanning multiple centuries, this variability in editorial maturation remains a background condition.

Altogether, these limitations point to a broader methodological constraint. Wikipedia's cultural entities are constructed knowledge objects shaped by the interaction of historiography and editorial practice. The structural and semantic regularities are therefore properties of this representation of the past. Their correspondence with established historical knowledge provides a useful validation signal, but it is not independent of the same traditions that inform Wikipedia itself, introducing a degree of circularity.

At the same time, this does not diminish the analytical value of the approach. Precisely because Wikipedia aggregates and formalises a vast body of historiographical knowledge, it offers a unique opportunity to study how cultural memory is structured, connected, and hierarchically organised at scale. The methods developed here make it possible to systematically interrogate these structures, to identify both dominant patterns and omissions, and to open new lines of inquiry into how knowledge about the past is collectively produced and represented.

In summary, the limitations outlined above qualify the scope of the findings presented in this thesis without undermining their substance. Together, the three contributions demonstrate that Wikipedia's cultural network encodes historically meaningful structural and semantic patterns at scale, that those patterns can be recovered through reproducible computational methods, and that this approach both corroborates patterns historians have documented through other means and surfaces quantitative regularities that were not previously visible at this scale.

4.4. Future Work

Several directions for future research follow from these limitations. A first direction concerns the identification of the nodes that occupy the core and the periphery within

each disciplinary cluster. The bimodal distribution of intra-cluster strength established that a core-periphery structure exists, but the question of which specific nodes constitute the core in each discipline requires a dedicated analysis. This would bring the present method into closer dialogue with debates in intellectual history about canon formation and the structural conditions of intellectual centrality.

A second direction concerns the generative mechanisms underlying the observed distributions. The bimodal-versus-power-law contrast between intra-cluster and inter-cluster connectivity is consistent with the coexistence of random-growth and preferential-attachment processes but confirming this dual mechanism would require explicit modelling. Building generative models that reproduce both distributions simultaneously, and testing whether they reproduce the empirical patterns observed across multiple periods, is a direction that would offer substantial theoretical rewards.

A third direction concerns the geographical and linguistic extension of the framework. The present analysis is restricted to the English version of Wikipedia. Applying it to other language editions would allow systematic comparison across cultural traditions. For instance, the relationships between religion and politics, or between scientific and literary cultures are all candidate domains in which the same analytical framework could be deployed.

A fourth direction concerns the extension of the data source. The method depends on a network with sufficient internal linking density, but it is not intrinsically tied to Wikipedia. Natural language processing tools could be used to extract comparable networks from other linked corpora, such as archival catalogues, bibliographic databases, or domain-specific knowledge bases. Comparing the cultural networks recovered from different sources for the same period would offer a way of testing the robustness of the findings reported here.

Bibliography

1. Langrock I, González-Bailón S (2022) The Gender Divide in Wikipedia: Quantifying and Assessing the Impact of Two Feminist Interventions. Journal of Communication jqac004. https://doi.org/10.1093/joc/jqac004

2. Tripodi F (2023) Ms. Categorized: Gender, notability, and inequality on Wikipedia. New Media & Society 25:1687–1707. https://doi.org/10.1177/14614448211023772

3. Lemieux ME, Zhang R, Tripodi F (2023) "Too Soon" to count? How gender and race cloud notability considerations on Wikipedia. Big Data & Society 10:20539517231165490. https://doi.org/10.1177/20539517231165490

4. Beytía P (2020) The Positioning Matters: Estimating Geographical Bias in the Multilingual Record of Biographies on Wikipedia. In: Companion Proceedings of the Web Conference 2020. ACM, Taipei Taiwan, pp 806–810

5. Graham M, Dittus M (2022) Geographies of digital exclusion: data and inequality. Pluto Press, London

6. Callahan ES, Herring SC (2011) Cultural bias in Wikipedia content on famous persons. J Am Soc Inf Sci 62:1899–1915. https://doi.org/10.1002/asi.21577

7. Miquel-Ribé M, Laniado D (2018) Wikipedia Culture Gap: Quantifying Content Imbalances Across 40 Language Editions. Front Phys 6:54. https://doi.org/10.3389/fphy.2018.00054

8. Johnson I, Lescak E (2022) Considerations for Multilingual Wikipedia Research

9. Eom Y-H, Aragón P, Laniado D, et al (2015) Interactions of Cultures and Top People of Wikipedia from Ranking of 24 Language Editions. PLOS ONE 10:e0114825. https://doi.org/10.1371/journal.pone.0114825

10. (2026) Wikipedia:Notability. Wikipedia

11. (2026) Wikipedia:Reliable sources. Wikipedia

12. (2026) Wikipedia:Independent sources. Wikipedia

13. Kourany JA (1997) HISTORY OF PHILOSOPHY: Disappearing Ink: Early Modern Women Philosophers and Their Fate in History. In: Philosophy in a Feminist Voice. Princeton University Press, pp 17–62

14. Kim S, Park S, Hale SA, et al (2016) Understanding Editing Behaviors in Multilingual Wikipedia. PLoS ONE 11:e0155305. https://doi.org/10.1371/journal.pone.0155305

Acknowledgments

In their book *Mapping Texts: Computational Text Analysis for the Social Sciences*, Dustin Stoltz and Marshall Taylor say that "whenever we encounter a text, an interconnected network of people, groups, and other texts are implicated". Inevitably, this thesis, beyond the infinite number of coding, thinking, and writing it required, also required innumerable conversations and inspiration from a vast network of people. Therefore, here I would like to thank all of these people and to apologise in advance if I am forgetting someone.

First of all, I would like to thank both my supervisors **Dr. Gustavo Ariel Schwartz** and **Dr. Juan Luis Suárez** for their help, guidance, and support over the last few years. They allowed me to work within very interdisciplinary environments and to some extent their scientific and humanistic way of thinking greatly influenced my approach. Muchas gracias!

Secondly, but most importantly, I would like to thank my family: **Miltiadis Agapitos, Giouli Chasapi,** and my brother **Panagiotis Agapitos**. Words cannot really describe how grateful I am to them for their support all these years. One thing that they taught me is that working as a team is the only way to solve every problem. Σας ευχαριστώ πάρα πολύ και ελπίζω να σας κάνω περήφανους!

This section would be empty if I did not mention the most important person in my life: **Dr. Zuzanna Lawera**. You are an inspiration to me; from how you think and solve problems to how beautifully you see life. You kept me sane throughout this process, and you fed my very competitive spirit by constantly beating me in Scrabble. Bardzo dziękuję!

Another group of people that I would like to thank **Dimitris Chasapis, Marianna Lepesioti, Aggelos Chasapis,** and **Romanos Chasapis**. They all taught me in their own

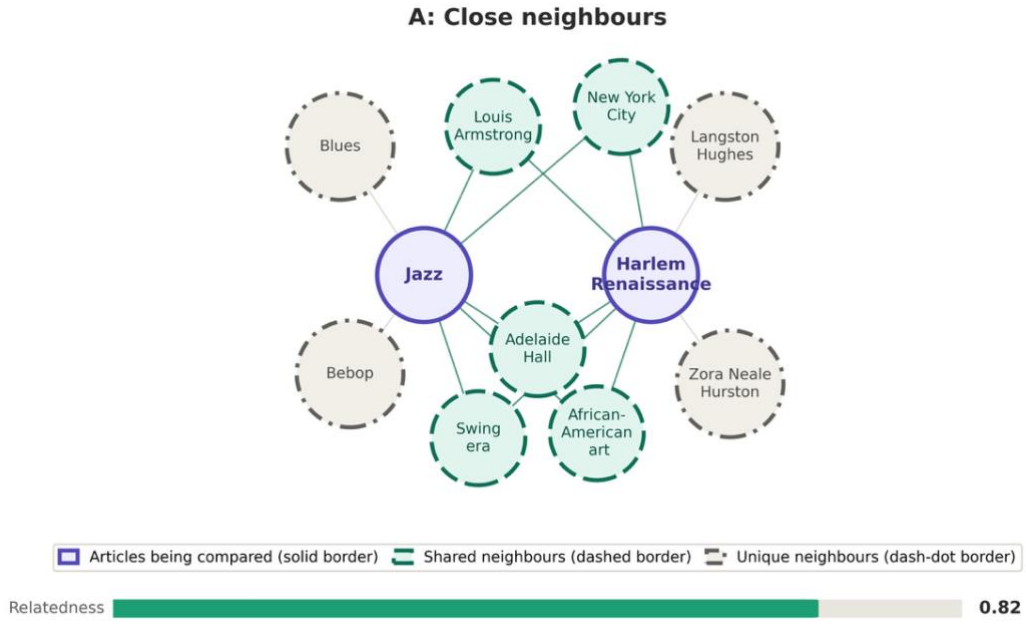
unique way to fight hard and not to back down no matter how difficult things can get. Σας ευχαριστώ πάρα πολύ!

My grandparents could not be missing from here either. After all, I owe them a great deal for much of what I have achieved in my life. **Ο παππούς Πασχάλης, η γιαγιά Πόπη, η γιαγιά Λούλα και ο παππούς Τάκης** ήταν, είναι και θα είναι πάντοτε παραδείγματα για μένα. Σας ευχαριστώ πάρα πολύ!

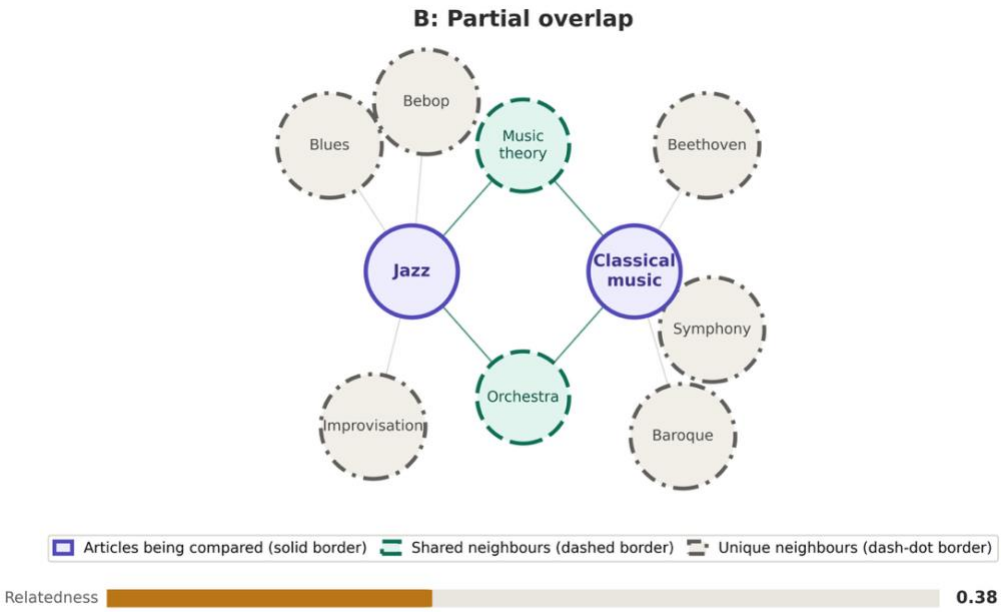
Appendix

A1. Period characterization

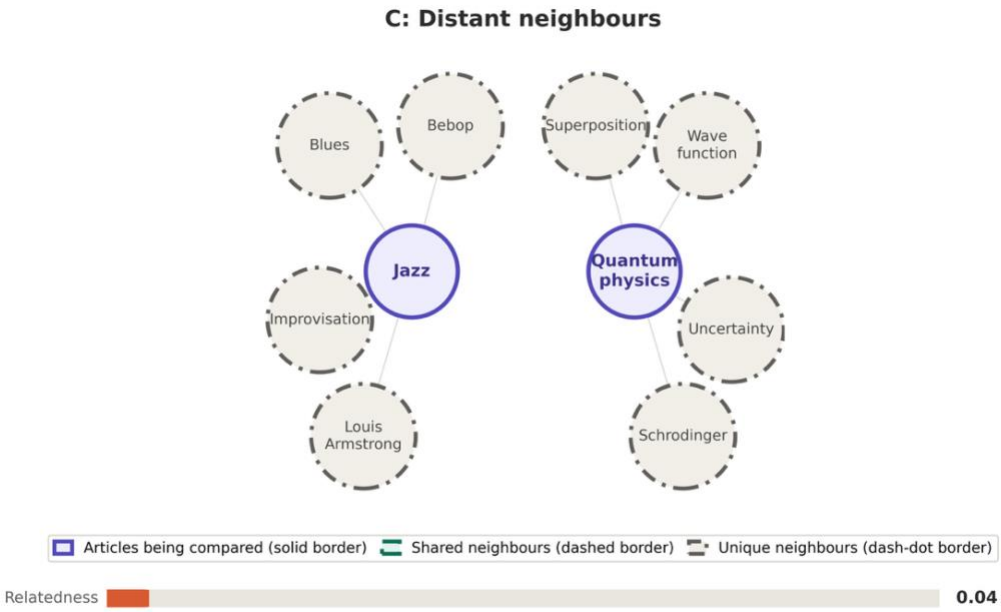
A1.1 Normalised Google Distance



Supplemental Figure 1 Fictional example to illustrate how structural relatedness between Wikipedia articles is measured by the Normalised Google Distance (NGD). The panel illustrates how the degree of shared neighbourhood between two articles determines their relatedness score. Purple nodes represent the two articles being compared; teal nodes represent their shared neighbours; grey nodes represent neighbours unique to one article. The relatedness bar at the bottom of each panel shows the corresponding score on a scale from 0 (no relatedness) to 1 (maximum relatedness). When two articles share a large proportion of their neighbours, as in the case of Jazz and the Harlem Renaissance, which are both connected to Louis Armstrong, Adelaide Hall, and New York City, the NGD assigns a high relatedness score (0.82), reflecting strong conceptual proximity despite the absence of a direct wikilink.



Supplemental Figure 2 Fictional example to illustrate how structural relatedness between Wikipedia articles is measured by the Normalised Google Distance (NGD). Each panel illustrates how the degree of shared neighbourhood between two articles determines their relatedness score. Purple nodes represent the two articles being compared; teal nodes represent their shared neighbours; grey nodes represent neighbours unique to one article. The relatedness bar at the bottom of each panel shows the corresponding score on a scale from 0 (no relatedness) to 1 (maximum relatedness). When the shared neighbourhood is modest relative to each article's total neighbourhood, as with Jazz and Classical music, which share only Music theory and Orchestra, the score drops to a moderate value (0.38), indicating partial but limited conceptual overlap.



Supplemental Figure 3 Fictional example to illustrate how structural relatedness between Wikipedia articles is measured by the Normalised Google Distance (NGD). Each panel illustrates how the degree of shared neighbourhood between two articles determines their relatedness score. Purple nodes represent the two articles being compared; teal nodes represent their shared neighbours; grey nodes represent neighbours unique to one article. The relatedness bar at the bottom of each panel shows the corresponding score on a scale from 0 (no

relatedness) to 1 (maximum relatedness). When two articles occupy entirely separate regions of the network and share no neighbours at all, as with Jazz and Quantum physics, the distance is effectively infinite and the relatedness score approaches zero (0.04), indicating the complete absence of any structural relationship.

A2. WikiTextGraph

A2.1. Page-level filter

Supplemental Table 1 Page titles excluded during the text preprocessing phase of the English Wikipedia corpus.

English (EN)				
Wiktionary:	Category:	Draft:	File:	List of
MediaWiki:	Module:	Template:	Wikipedia:	Index of
Help:	Portal:	Image:	Draft:	(disambiguation)

A2.2. Sections filter

Supplemental Table 2 Section where WikiTextGraph will stop the extraction of the text. The sections below signify blocks that usually contain external links and not wikilinks.

English (en)			
See also	Publications	References	Notes
Footnotes	External links	Further Reading	

A3. Temporal and Semantic Patterns In Wikipedia's Network of Biographies

A3.1. Fields of Human Activity (FHAs)

The final list of FHAs along with a brief description of what they include are:

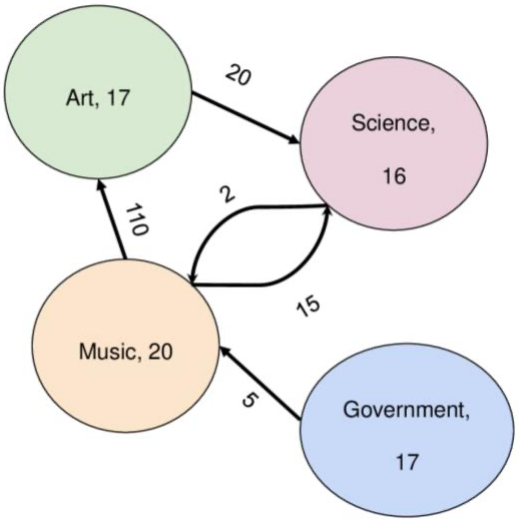
1. **Sports:** Athletes of all types, such as track and field competitors, runners, and F1 drivers.
2. **Government & Law:** Encompasses politicians (e.g., senators, royalty) and those in Law Enforcement (e.g., police officers, jury members, attorneys).
3. **Dramaturgical Arts:** Covers mostly actors and actresses.
4. **Applied Arts:** Features individuals involved in any form of applied art, such as sculptors and painters.
5. **Music:** Includes musicians like singers, front-persons, drummers, and conductors.
6. **Literature:** Encompasses those connected to literature, such as writers, authors, essayists, and translators.
7. **Humanities & Philosophy:** Contains individuals within the broader fields of humanities and philosophy, including philosophers, educators, teachers, and professors.
8. **Business:** Includes people involved in business, such as entrepreneurs, businesspeople, bankers, and merchants.
9. **Science:** Covers individuals in scientific fields, such as physicists, engineers,

chemists, biologists, and astronauts.

10. **Religion:** Includes people related to religion, such as priests and religious leaders.
11. **Media:** Encompasses individuals in media, such as journalists, radio hosts, and YouTubers.
12. **Military:** Covers individuals associated with military or paramilitary organizations, including generals, soldiers, and paramilitary operatives.
13. **Healthcare & Medicine:** Covers individuals in the healthcare and medical fields, such as doctors and paramedics.

A3.2. Weight Normalisation

The following example illustrates how connection counts are normalized under each method. Consider a **hypothetical** directed weighted network *G* (**Supplemental Figure 4**). In this network, nodes represent clusters defined by a century and an FHA. The edges connecting these nodes indicate a relationship between them. This relationship is expressed through the initial weight assigned to each edge corresponds to the number of connections originating from one node and leading to another. For example, "(Art, 17) → (Science, 16), 20" indicates the existence of 20 links directed from artists in the 17th century CE to scientists in the 16th century CE.



Supplemental Figure 4 Dummy graph as an example to explain the different types of normalisations. The colour of each node corresponds to the FHA and the century. The edges are being weighted based on the number of links that each (FHA, century) sends to another (FHA, century).

Out-degree normalisation

For the first scenario of normalizing based on the outgoing links, we start by building the raw weight matrix of the network $G: W = [w_{ij}]$ which looks like this (**Supplemental Table 3**):

Supplemental Table 3 Raw weight matrix for outgoing-link normalisation. The table reports the unnormalized adjacency weights w_{ij} of network G, where each cell indicates the number of outgoing links from biographies in a given source field–century group to biographies in a corresponding target group. Source groups are listed in rows and target groups in columns.

		Target			
		Art, 17	Science, 16	Music, 20	Government, 17
	Art, 17	0	20	0	0
	Science, 16	0	0	2	0

Source	Music, 20	110	15	0	0
	Government, 17	0	0	5	0

At this stage we aim to normalize based on the outgoing links (row-wise). Thus, we do:

$$w^{+}{}_{ij} = \frac{w_{ij}}{\Sigma_{j=1}^{N} w_{ij}}$$

Where:

- w_{ij} is the raw number of connections from node i to node j.
- $\Sigma_{j=1}^{N} w_{ij}$ is the total sum of connections **leaving** from node i.
- $w^{+}{}_{ij}$ is the normalized weight based on the outgoing links/ out-degree.

The resulting table (**Supplemental Table 4**) looks as follows:

Supplemental Table 4 Row-wise normalisation of outgoing link weights. The table presents the adjacency matrix after normalizing each row by the total number of outgoing links from the corresponding source group. Each entry shows the proportion of links from a given source field-century group to each target group, yielding normalized weights that sum to 1 within each row.

		Target			
		Art, 17	Science, 16	Music, 20	Government, 17
	Art, 17	0	20/20 = **1.00**	0	0

Source	Science, 16	0	0	2/2 = **1.00**	0
	Music, 20	110/125 = **0.88**	15/125 = **0.12**	0	0
	Government, 17	0	0	5/5 = **1.00**	0

For any cluster Source in any given century, out-degree normalisation tells us: "Out of all the links that cluster Source sends out, what percentage goes to cluster Target?" This means, for each source category, we look at all the outgoing links and calculate what fraction of them are directed to each possible target category. The fraction highlights the relative importance of different target categories from the perspective of the source. To illustrate it more clearly, we will give an example on how to interpret the values. Suppose that we study the (Art, 17 - Science, 16) relationship from Table SM2.

- Art, 17 → Science, 16: 20 / 20 = 1.0
- 20 edges leave from Art, 17 and they end up in Science, 16.
- The total number of outgoing links (deg^+) of Art, 17 is 20.
- Interpretation: 100% of the edges that leave from Art, 17 end up to Science, 16.

In summary, out-degree-based normalisation gives the proportion of the source's outgoing links that are directed to each specific target.

In-degree normalisation

For the second scenario, which is the normalisation based on the incoming links, consider again the same **hypothetical** directed weighted network. As in the previous example, the same raw weight matrix: $W = [w_{ij}]$ which look as follows (**Supplemental Table 5**):

Supplemental Table 5 Raw weight matrix used for incoming-link normalisation. The table displays the unnormalized adjacency weights w_{ij} for the hypothetical directed network used in the incoming-link normalisation scenario. Each cell reports the number of links received by field-century group from a given source group, forming the basis for subsequent column-wise (incoming) normalisation.

		Target			
		Art, 17	Science, 16	Music, 20	Government, 17
Source	Art, 17	0	20	0	0
	Science, 16	0	0	2	0
	Music, 20	110	15	0	0
	Government, 17	0	0	5	0

Normalisation based on the in-coming links necessitates performing the previously executed procedure but now in a column-wise manner. Thus:

$$w^{-}_{ij} = \frac{w_{ij}}{\Sigma_{j=1}^{N} w_{ij}}$$

Where:

- w_{ij} is the raw number of connections from node *i* to node *j*.
- $\Sigma_{j=1}^{N} w_{ij}$ is the total sum of connections **coming** to node *j*.
- w^{-}_{ij} is the normalized weight based on the incoming links/ in-degree.

The weight in this case will be normalized as follows (**Supplemental Table 6**):

Supplemental Table 6 The table reports the adjacency matrix after normalizing each column by the total number of incoming links received by the corresponding target group. Each entry represents the proportion of links directed toward a target field–century group that originate from a given source group. This normalisation highlights the relative contribution of each source to the incoming link profile of each target.

		Target			
		Art, 17	Science, 16	Music, 20	Government, 17
Source	Art, 17	0	20/35 = **0.57**	0	0
	Science, 16	0	0	2/7 = **0.28**	0
	Music, 20	110/110 = **1.00**	15/35 = **0.43**	0	0
	Government, 17	0	0	5/7 = **0.72**	0

For any cluster Target (FHA, century), in-degree-based normalisation answers the question: "Out of all the links that cluster Target receives, what percentage comes from cluster Source?" In other words, for each target cluster, we look at its incoming links

and calculate what fraction originated from each possible source cluster. This fraction helps us understand the relative importance of source clusters as the main contributors of links to other target clusters. As in the out-degree-based normalisation we provide an example to help the readers understand better the interpretation of the values after applying in-degree-based normalisation. Assume that we want to study the (Art, 17 - Science, 16) relationship (**Supplemental Table 6**).

- Art, 17 → Science, 16: 20 / 35 = 0.57
- 20 edges leave from Art, 17 and end up in Science, 16.
- The total incoming links to Science, 16 (deg^-) is 35.
- Interpretation: 57% of the edges that come Science, 16 are from Art, 17.

In summary, in-degree-based normalisation gives the proportion of the target's incoming links that are coming from each specific source.